

Galaxy Formation: an Uncertainty Analysis

Ian Vernon¹, Michael Goldstein¹, Richard Bower²

Department of Mathematical Sciences¹

Department of Physics²

Durham University

Science Laboratories

South Rd

DURHAM DH1 3LE

UNITED KINGDOM

`i.r.vernon@durham.ac.uk`

July 25, 2009

Contents

1	Introduction	4
2	A universe full of galaxies	6
2.1	Understanding our place in the cosmos	7
2.2	Galaxy Formation - a beginners Guide	8
2.3	Modeling Galaxy Formation	10
2.4	The Galform model	11
3	Uncertainty Analysis for Computer Simulators.	12
3.1	Uncertainty in complex models	12
3.2	Linking the simulator with the system	14
3.3	Bayes Linear Analysis	16
3.4	Emulation	17
3.5	History Matching	19
4	The Galform Model	21
4.1	The Dark Matter Simulation	22
4.2	The Galform Model	22
4.3	Galform: Physical Details	23
4.4	Inputs	26
4.5	Outputs	26
5	First Wave Analysis	28
5.1	The Wave 1 Emulator	28
5.1.1	General Designs for Computer Model Experiments	28
5.1.2	The Wave 1 Design	28
5.1.3	Emulator Construction	32
5.1.4	The Wave 1 Emulator	34
5.2	Diagnostics	36
5.2.1	Residuals	37
5.2.2	Emulator Prediction Diagnostics	38
6	Quantification of Uncertainty	40
6.1	Model Discrepancy	41
6.1.1	Uncertainty Due to Inactive Variables: Φ_{IA}	41
6.1.2	Dark Matter Uncertainty: Φ_{DM}	43
6.1.3	Full Galform Model Discrepancy: Φ_E	44
6.2	Observational Errors	47

7	First Wave History Match	49
7.1	Implausibility Measures	49
7.2	History Matching via Implausibility	51
7.3	Refocusing	55
8	Analysis of Waves 2 - 4	57
9	Results of Wave 4 and 5 - Visualisation	57
10	Conclusion	57
A	Wave 1 Polynomials	59

1 Introduction

Current theories of cosmology suggest that the Universe began in a hot, dense state approximately 13 billion years ago, and that it has been expanding rapidly ever since. However, observations of galaxies imply that there must exist far more matter in the Universe than the visible matter that makes up stars, planets and us. This is referred to as ‘Dark Matter’ and understanding its nature and role in the evolution of galaxies is one of the most important problems in modern cosmology. The Galform group, based at the Institute of Computational Cosmology, Durham University, is the world leading group in the study of Galaxy Formation in the presence of Dark Matter. Over the last 13 years, they have developed a detailed computer model, known as Galform, which simulates the creation and evolution of approximately one million galaxies from the beginning of the Universe until the present day. The simulation produces various physical features of each of the galaxies which can be compared to observed galaxy survey data.

The Galform model requires many input parameters to be specified in order to run the simulation. It is therefore necessary to explore the input parameter space and find the set of all input configurations that give rise to acceptable matches between model output and observed data. As the model run time is significant, this is a challenging task. Further, even to assess what constitutes an acceptable match, we must consider all of the uncertainties that are involved in the comparison between model and reality, including input parameter uncertainty, function uncertainty, observational error, forcing function uncertainty and structural uncertainty. Such a detailed level of uncertainty quantification has never been attempted for a cosmological model of this size and complexity.

This case study describes a collaboration between members of the Statistics group and the Galform group, at Durham, to carry out such an uncertainty analysis for Galform. Our aim is to identify all choices of input parameters that generate consistent physical models in the sense that they would yield sufficiently good matches to certain important features of observational data, when we have taken into account all relevant sources of uncertainty. In particular, it is of fundamental interest to know whether this set of acceptable inputs is non-empty.

In order to treat all uncertainties in a consistent and unified manner, we use general techniques related to the Bayesian treatment of uncertainty for computer models for large scale physical systems. In addition to the uncertainty associated with the Galform function itself, we elicit all of the other sources of uncertainty which must be addressed in order to make meaningful comparisons between Galform output and observational measurements, in particular, making expert assessments for the structural uncertainty which arises due to the inherent limitations of the physical model.

Our approach is based on the construction of an emulator for Galform, this being a stochastic function that represents our beliefs about the behavior of the simulator. We use the emulator and the model uncertainties to define implausibility measures over the input parameter space for Galform, based on a Bayes Linear analysis. High values of the implausibility measures suggest that we should consider that it is very unlikely that an acceptable match to the chosen observational features would be obtained by evaluating

the model at the corresponding input values, and hence we can exclude regions of input space by imposing cutoffs on our implausibility measures. We proceed iteratively, making function evaluations over the full range of the input space, emulating Galform over this space, using implausibility measures to remove a part of the space, making a further collection of evaluations of Galform in the reduced space, re-emulating within the reduced space, re-evaluating our implausibility measures over this subspace and therefore removing a further portion of the space and continuing in this fashion. We have performed this cycle four times, in each case making a substantial further reduction to the allowable input space. Our final stage was to make a further set of runs to check that we did indeed have a large number of acceptable matches between Galform output and observations over a range of input parameter choices within the final reduced space.

Visualising the outcome of this process is a non-trivial task. The acceptable region of input space is a complex shape in high dimension. Although the emulators are fast to evaluate, they still cannot give sufficiently detailed coverage of the full volume. We have therefore developed fast emulation techniques specifically targeted at producing lower dimensional visualisations of higher dimensional objects, leading to novel dynamic 2 and 3 dimensional projections of the acceptable input region. This is a significant contribution toward understanding the Galform model, as previously no knowledge of the shape and extent of the acceptable region of input space existed. Further, the previous best matches to the primary data set of interest were not compatible with other secondary, but important, observational data sets. Our analysis demonstrates that, by making realistic assessments of structural uncertainty, we are indeed able to simultaneously match data sets that were previously thought to be incompatible, contradicting authors who suggested the Universe is ‘anti-hierarchical’ and such a match impossible. Thus this work should be viewed as supporting the hypothesis that galaxies formed in the presence of large amounts of Dark Matter, and in particular via hierarchical merging.

This collaboration began in an informal fashion. Members of the statistics group were interested in applying various techniques that they had developed for the analysis of large scale computer models, aspects of which were reported in a previous Case studies meeting at Pittsburgh [6]. The Galform group offered the use of their model and some of their computing facilities. Over time, and after many discussions and preliminary explorations, it became clear that such an analysis was a useful tool for understanding various scientific issues related to the model, and merited a serious collaborative effort to pursue these questions. This account is a description of the results of the collaboration, described more or less as it has evolved.

The Case Study paper is structured as follows. In section 2 we discuss the physical motivation for the study of galaxy evolution and give a general description of the Galform model. Section 3 describes the Computer Model methodology that we will employ, and highlights all the relevant uncertainties that must be considered. The details of the Galform Model necessary for an uncertainty analysis are given in Section 4, along with further physical description, and in section 5 we describe the construction of the Wave 1 emulator. In section 6 we assess all remaining uncertainties relevant to the analysis and in section 7 we perform

the first iteration of the History Matching process. Later sections deal with the second, third and final iterations, describe the various visualisation techniques developed to interpret the results, and conclude with discussions regarding the physical insight gained through such an analysis.

2 A universe full of galaxies

The night sky is full of stars. Yet the stars that are visible to the human eye are only an unimaginably tiny fraction of the stars in the universe as a whole. Equipped with telescopes, we discover that at great distances beyond our own galaxy lie millions of millions of other galaxies, each with their own populations of stars. As a philosophical idea, these “island universes” date back to Kant (1755), but experimental determination of the great distance to external galaxies has only been established surprisingly recently (Hubble 1929).

Galaxies come in great variety of shapes and forms. Our own Milky Way galaxy is one of the larger spiral type galaxies. Spiral galaxies are dominated by a flat disk of stars, often with prominent spiral arms. The Milk Way’s stellar disk can be seen in the night sky as the prominent band of stars that gives us its name. In addition to stars, spiral galaxies contain significant amounts of gas and dust that can be seen to fuel the birth of further generations of stars. Although spiral galaxies are the most numerous, the most massive galaxies have a very different appearance. Largely devoid of gas and dust, they have a 3-dimensional ellipsoidal appearance. Such elliptical galaxies are generally found in dense associations of galaxies referred to as galaxy clusters. Hubble (1936) established a well defined system for classifying the appearance of galaxies, commonly referred to as Hubble’s “tuning fork”, in which he notes a continuum of galaxy properties with the central parts of spiral galaxies becoming more and more dominated by an elliptical-like bulge. He also noted that many smaller galaxies were too disturbed to fit into this classification system, denoting such galaxies “irregular”. The most spheroid dominated galaxies are referred to as “early-type”, while the most disk dominated (and irregular) systems are referred to as “late type”. Although based primarily on the significance of the bulge component and the prominence of spiral arms, the Hubble sequence is also closely connected to a galaxy’s star formation rate and its colour, Late-type (spiral and irregular) galaxies being the most prodigously active and the early-types showing little or no on-going star formation. Although Hubble used the terms “early” and “late”, he did not intend it to describe an evolutionary sequence, and indeed current theories suggest that “early” type galaxies are formed from mergers between (and instabilities in) late type galaxies.

With modern telescopes, it has become possible to study galaxies at greater and greater distances from earth. Because of the finite speed of light, such distant galaxies are seen when the universe was younger. Astronomers can use this time delay to observe the build up and formation of galaxies. The most distant galaxies identified to date are seen only 10^9 yr after the big bang, when the universe was less than $1/10^{\text{th}}$ of its current age. These observations have revealed some, at first sight, puzzling results. The “natural” sequence for the formation of galaxies is through a process of hierarchical aggregation: small galaxies form early in the

history of the universe, building larger and larger galaxies through gravitational collapse. This picture is a natural consequence of the Cold Dark Matter model that describes the large scale properties of the COSMOS well. The picture is however at odds with observational studies that find a large proportion of the most massive galaxies are present quite early in the history of the universe. Explaining the tension between the prima-face theoretical expectation and the observational evidence was one of the key motivation for developing the theoretical model that is discussed below.

2.1 Understanding our place in the cosmos

The aim of galaxy formation studies is to understand why the universe appears as it does. We wish to explain the characteristic properties of galaxies, such as their distribution of luminosities, sizes and ages. In doing so, we are understanding what makes the universe tick. This purpose is part of an age old quest to understand our origins in the deepest sense. It is obvious that, without stars, there could be no life. Yet it is equally true that without the large accumulations of stars that we know as galaxies we could not exist. Beyond hydrogen, the key elements that we are made of (such as, Carbon, Oxygen, Nitrogen, Iron) are only abundant because our sun was created from the leftover remnants of previous stars. Without sufficient earlier generations of stars, or the galaxy’s gravity to capture and recycle these elements, it is unlikely that the formation of the sun would be accompanied by the formation of rocky planets like the Earth and Mars. Galaxies are key to our existence. Understanding the creation of galaxies is key to our ability to observe the cosmos.

As we will describe below, the present problem is not so much to understand why galaxies form, but to understand why they are relatively few and far between. By understanding this, we hope also to explain why galaxy formation appears to proceed very differently to that expected in the simplest theories. The basic ingredients have been in place for some time (the force of gravity and radiative cooling of baryonic matter), but we are only now beginning to understand how the formation of galaxies is regulated. The surprising result is that the black holes (the densest objects in the universe) appear to play a key role in this.

The aims of the field go beyond understanding the formation of galaxies alone. Firstly, we apply this understanding to better constrain cosmological parameters. A key application of the galaxy formation models we describe below is to generate “mock” galaxy catalogues — catalogues of simulated galaxies as they would appear when observed by existing or proposed telescopes. Such catalogues are essential for understanding the biases in observational surveys. For example, current BAO (Baryon Accoustic Oscilation) surveys aim to measure the expansion of the universe using the imprint of the horizon scale at matter-radiation decoupling. Since galaxies are used as tracers of this structure, it is essential to quantify the interplay between their observational selection and the underlying matter power spectrum. The measurement of this horizon scale provides a key measurement of the geometry of the universe and the results give strong evidence for the existence of “Dark Energy”, or a significant vacuum energy density.

More speculatively, an accurate understanding of galaxy formation would allow us to address illusive anthropic issues, such as the viability of life in other universes where physical

constants have different values to those in our own. Of key philosophical interest is the interplay between the loss of entropy (and increase of order) due to radiative cooling, and its increase due to gravitational collapse. Seemingly the universe seems pre-programmed to generate sustainable life out of the disorder of the big bang.

2.2 Galaxy Formation - a beginners Guide

So how do galaxies form? Why is the universe filled with such objects? In principle, it is a straightforward consequence of the dominance of the gravitational force. Since all matter makes a positive contribution to the gravitational force, the clumping of the universe's mass is a run away process. As the condensations of matter become denser, they become more effective at attracting matter from around them. As a result, small quantum fluctuations early in the history of the universe are strongly amplified, building larger and larger density concentrations. These concentrations are referred to as haloes.

In a strange twist, the observational evidence shows that most of this mass, however, is not normal matter. Recent measurements have shown that the matter content of the universe is dominated by "Cold Dark Matter" (CDM): massive particles that interact very weakly with the normal, "baryonic" matter that you and I are made from (principally protons, neutrons, electrons and neutrinos). Current experiments at the LHC may demonstrate that the CDM particles are a fundamental consequence of super-symmetric extensions of the standard model of particle physics. These dark matter particles dominate the large scale dynamics of the universe and cause the collapse of local perturbations of excess density. Such collapsed regions are referred to as dark matter "haloes". They are the sites in which galaxies may form if the astrophysical conditions are favourable. For our purposes, this twist is convenient: we may think separately about the purely gravitational processes (that lead to the creation of dark matter haloes) and the astrophysical processes (that create luminous stars out of the normal baryonic matter). This separability is a fundamental idea that we will exploit below. Recent observations have, however, shown that an additional vacuum energy contribution is required to match the growth and clustering of dark matter haloes. Essentially this amounts to adding an additional long-range acceleration to the gravitational equations. Its physical origin is very poorly understood at present, but its impact on the formation of dark matter haloes can be accurately described.

Nevertheless, the CDM particles (together with the big bang and the vacuum energy) only explain the collapse and growth of the gravitating dark matter haloes. To describe the formation of the luminous galaxies, we must turn to the astrophysics of the baryonic matter. Our everyday experience tells us what happens. As the baryons are pulled together by the collapse of the dark matter halo, they heat up. This generates pressure which resists further compression of the baryonic gas. Galaxies form because the baryonic gas can radiate this thermal energy and cool. As a result, the baryons lose pressure and contract within the halo. However, the cooling rate increases strongly as the density increases, and the contraction only serves to accelerate the cooling until a run-away situation is reached. The contraction of the baryons is only stopped by the conservation of angular momentum and the baryons form thin, cold spinning disk of gas. Although we do not fully understand how the

gas condenses further into stars, empirical measurements show that the rate of formation of stars is proportional to the surface density of gas; for current theoretical models this empirical calibration is entirely sufficient. In this scenario, small haloes (where the initial gas temperature is low) are able to cool almost all their baryonic component into stars [indeed the gas may never actually be heated in the collapse], while in the largest haloes, the initial gas temperature is so high that the cooling timescale is longer than the age of the universe.

This scenario is simple, but does not reflect the universe we live in. The fraction of the baryonic material that is observed to form into stars (or cold gas) is rather small, only about 10% of the total baryonic content of the universe. The origin of this discrepancy is one of the principle cosmological puzzles. Astronomers appeal to “feedback” to resolve the discrepancy: somehow the formation of stars must inject energy that prevents further gas cooling. One of the key aims of the GALFORM project is to identify the feedback schemes that are needed to account for the observed universe. In small galaxies, this has long been understood (at least in principle). At their death, massive stars undergo supernovae: energetic explosions that are capable of driving gas out of the galaxy and re-heating to the initial temperature. This process is weak in our galaxy because the gravitational field is so strong, but in low mass galaxies it may drastically reduce the overall efficiency of star formation.

The strength and importance of feedback is best assessed by comparing the observed galaxy luminosity function (the numbers of galaxies in a given luminosity bin per unit volume) with the mass function (the number of a given mass per unit volume) of dark matter haloes. If star formation were uniformly efficient (and only one galaxy was formed in each halo) there would be a constant offset between the two functions. A comparison shows that they differ dramatically in shape: the dark matter mass function has far more small haloes than are observed to host dwarf galaxies in the universe. Supernova feedback explains the discrepancy between the observed functions at the faint end by flattening the predicted relation due to the inefficiency of star formation in small galaxies.

Unfortunately, realistic calculations show that while this simple fix may solve the problem with faint galaxies, it leads to a bigger problem for the most massive galaxies. Essentially, because we have reduced the fraction of baryons locked up in small objects, there is more material left over to form greater numbers of stars in the most massive galaxies. The models predict the formation of far too many galaxies much larger than the Milky Way. Various solutions have been proposed. One possibility is energetic “superwinds” that blast material out of young, forming galaxies; but the current front runner is a form of feedback associated with the slow accretion of gas on to black holes.

This form of “AGN” feedback is at first sight rather exotic. Black holes are the smallest objects in the universe, their size (measured as their Schwarzschild radius) is only 1.5×10^8 km. It is surprising that an object so small can heat a volume with radius 10^{11} times larger. Yet this is just what is observed in clusters of galaxies. Clusters are gravitationally bound systems containing 1000s of galaxies and 10^{15} solar masses of (largely) dark matter. Gas at the centers of these systems is dense enough that it should cool, promoting the formation of stars in the central object. Yet, no such cooling is observed. Instead these systems host a powerful radio galaxy — a galaxy with a central black hole (or AGN) that is the

source of a jet of magnetised high energy plasma. Recent observations have shown that the energy content of the plasma is very high - high enough to push the cooling baryons aside generating huge bubbles in the intra-cluster gas. Although the details are not yet clear, these jets are capable of replacing the energy that is lost as cooling, keeping the central gas hot and starving the central galaxy of fuel for star formation. The frequency of the discovery of such objects is also remarkable - they seem to occur everywhere the runaway cooling process would generate a problem.

Clearly this remarkable process is key to fully understanding of the formation of galaxies. It is now widely accepted that it provides an essential ingredient for models that explain the formation of galaxies.

2.3 Modeling Galaxy Formation

There are essentially two approaches to modelling the formation of galaxies. They are often seen as adversarial, but, in practice, they are highly complementary. These are usually referred to as “numerical simulation” and “semi-analytic modeling”.

The idea of “numerical simulation” is simple and direct. A powerful computer, is programmed with the fundamental physical equations that describe the growth of fluctuations of dark matter, the hydrodynamical response of the intergalactic gas and its loss of energy through key atomic cooling processes. Such a system of equations can be accurately tracked and a mathematically stable solution found. However, as we have described above, the equations are missing some key components of galaxy formation physics and, if left to themselves, massively over-produce the abundance of stars. Unfortunately, such codes have no hope of directly following the formation of stars or the winds they may generate at their death. They are many more orders of magnitude from being able to track the formation of black holes or the processes that generate the jets that we think regulate the formation of bright galaxies.

“Semi-analytic modeling” represents the alternative approach. Rather than tackling the whole problem in a single numerical integration, we break it down into its separate components. Of course, we must make some level of approximation by doing this, but we hope to create a model that encompasses the main physical processes with a minimum of complexity. For example, one component of the model is the growth and merging of dark matter haloes. This can be computed through an analytic approximation or by running a numerical calculation that only includes the force of gravity. In terms of the behaviour of the dark matter, this approximation is extremely good. We must then add components to describe such features as the collapse and cooling of gas; the formation of stars; the growth of black holes; merging of galaxies; the feedback effect of supernova explosions and jets from black holes, and then link them together through a network of interactions. Adding further components complicates the model but may improve its physical realism and ability to match the data. Each component is based on the results of a targeted set of simulations - or, failing this, on physically plausible scaling relations. In many cases, however, the physical process is not completely understood or characterised: to cope with this we introduce a number of parameters to account for this uncertainty. The values may be quite tightly constrained, or may have large uncertainty depending on the physical complexity that they capture. The

result is a network of equations (or algorithms) whose behaviour is driven by the underlying growth and merging of the dark matter haloes, and whose response is governed by a number of adjustable parameters. Often the response of the system is counter intuitive - for example, adjusting the star formation rate timescale has little effect on the mass of stars formed, tending instead to simply alter the mass of the cold gas reservoir that is waiting to form stars.

Because of the intrinsic complexity of the galaxy formation problem, “semi-analytic models” currently offer the best avenue for progress. Prior to the introduction of AGN feedback, the models struggled to reproduce realistic galaxy properties. For the reasons described previously, the models struggled to match the observed shape of the luminosity function, as well as a number of the other observable properties of galaxies. The current generation of semi-analytic models use AGN jets to solve the problem of the over formation of bright galaxies, although different authors use very different approaches to including the new AGN feedback physics. The GALFORM model was one of the first models to include this process [2], showing that the AGN jet feedback gave the model the freedom to match the observed luminosity function. Although the model was adjusted to work on using local data, it was found that the model also provided a good description of the luminosity function throughout cosmic history: accounting for the “anti-hierarchical” observations of a significant population of early, massive galaxies.

A galaxy model that has been tuned to match the local luminosity function, also makes “predictions” for other properties of galaxies. For example, the colour distribution, sizes and rotation velocities of galaxies can also be compared to observational measurement. The GALFORM model does moderately well in these comparisons — making predictions of the correct magnitude — but the match to the data is far from perfect. Clearly an important question is whether it is possible to adapt the parameters of the model so that the fit to the luminosity functions remains comparably good, but the match to additional datasets is improved. Failure to improve the model in this way would suggest that the code is missing some key physical processes that are important for galaxy formation.

2.4 The Galform model

The GALFORM code is a world-leading semi-analytic galaxy formation model. The code separates the physical processes involved in galaxy formation into modules. The principle modules track:

1. the gravitational collapse and build-up of dark matter haloes;
2. the cooling and accretion of gas; the formation of stars, stellar evolution and “feedback” from supernova explosions;
3. galaxy mergers and instabilities in stellar disks;
4. the formation of black holes and the associated feedback;
5. the effects arising from re-ionisation of the universe by the ultra-violet radiation field.

The computer code for each of these sections implements astrophysically motivated algorithms, each process drawing on the inputs provided by each of the other modules. The modules link together to form a network of non-linear equations that are integrated in time to trace the evolving properties of the galaxy population. The coding of each individual module is quite complex. In total the model uses over 50,000 lines of computer code. Further details of the modules are described in a later section. [1] presents a suitable introduction to the internal workings of the code.

Each module has associated parameters. These define the working of each module. For example, they specify the rate at which cold gas is converted into stars; or the energy generated in supernova feedback and its dependence on galaxy mass. In order to run the code, the astrophysicist must specify values for each of these parameters. Some parameters are quite well defined by numerical experiments or targetted observational data, but others are highly uncertain. Conventionally, the astrophysicist makes an educated guess at plausible values of the parameters, and then adapts the values to converge slowly on an acceptable solution. This process is slow, relies heavily on the modeler's intuition and is readily trapped into local solutions. Clearly this is an area which could be hugely improved by applying systematic methods for uncertainty analysis to explore the parameter space, and this provides the motivation for the current Case Study.

3 Uncertainty Analysis for Computer Simulators.

3.1 Uncertainty in complex models

Our aim in this case study is to identify that region of the input space of the Galform simulator for which certain aspects of Galform output match closely to measurements that have been made in the observable universe. As such, this study falls within the general area of the analysis of uncertainty arising when we study complex physical systems by means of mathematical models typically implemented as computer simulators.

The general version of the problem is as follows. We have a computer simulator f which takes as input the vector x , which represents certain physical properties of a system of interest. The simulator output vector, $f(x)$, corresponds to certain aspects of the behaviour of the system. For a given choice of inputs, this behaviour is determined, in principle, by a series of equations embodying all of the relevant theoretical knowledge relating system properties to system behaviour.

This approach is common to many areas of science. Climate simulators are used to study climate, reservoir simulators are used to study reservoirs, flood simulators to study flooding and, in our case, a universe simulator is used to study the universe. The reason that we can talk of an emergent methodology to address each problem is that, despite the enormous differences between each of the individual models, all such problems of physical modelling will need to confront a similar collection of basic uncertainties.

[1] **Parameter uncertainty.** We do not know the appropriate values of the inputs to the simulator. In some cases, we may not even know whether there is any appropriate

choice for the inputs. Galform is a case in point. If we have misrepresented the underlying physics, for example if it turns out that the current view of the role of Dark Matter is not supported by the weight of observational evidence, then the basic meaning of the model and the interpretation of the parameters will be called into question. In particular, were we to discover that there were no choices of inputs for which Galform output matched observations in our universe, then that might provide part of the evidence which would call the current account of cosmology into question. Our aim in this study is to determine whether there are any acceptable choices of input values for Galform, and, if so, to describe the collection of such possible choices.

[2] Simulator uncertainty. For any choice of inputs, x , the output $f(x)$ is a deterministic computer function. However, many computer simulators are very expensive, in time and resources, to evaluate, for any choice of inputs. In practice, it is appropriate to consider that the output values of such a simulator are unknown except at the input choices at which the simulator has been evaluated. An important stage in the analysis, therefore, is the construction of a statistical representation or **emulator** for the simulator. The emulator represents our uncertainty about the value of the function at each possible input choice, and therefore acts both as an approximation to the function and as an assessment of the uncertainty introduced by the approximation. Much of the literature on **computer experiments** is concerned with efficient choices of designs to determine a collection of evaluations of the function which will allow us to build an emulator which is effective in performing certain tasks associated with the function, for example finding the maximum value of some combination of the outputs; see for example [7, 16, 17]. For our Galform investigations, we have been fortunate in being able to make a large number of evaluations of the simulator. Even so, emulation has proved to be a key step in extending our uncertainty description from the function evaluations to the remainder of the input space.

[3] Structural uncertainty. This refers to the fundamental problem that, however carefully we have constructed our model, there will always be a difference between the system and the simulator. A climate model will never be the same as climate, and nor would we expect it to be. Inevitably, there will be simplifications in the physics, based on features that are too complicated for us to include, features that we do not know that we should include, mismatches between the scales on which the model and the system operate, and simplifications and approximations in solving the equations determining the system. Often, understanding this structural uncertainty will be one of the most challenging aspects of the analysis. The interweaving of the emulation technology developed within the computer experiment literature and the careful consideration of structural uncertainty is, in our view, the driving force for this new area of statistical methodology. We shall pay close attention to structural uncertainty in the current study, as our judgements as to the quality of match to the observable universe that we should demand from our Galform evaluations depends critically on the fidelity to observation which we expect the simulator to be capable of achieving.

[4] Observational error. This type of uncertainty arises when we consider the match of our model to system observations. Measurement errors are familiar, in principle, to

statisticians. However, it is often the case, in complex physical systems, that the observations are themselves somewhat indirect, being assessed on the basis of extensive preprocessing based on various additional theoretical constructs. Further, the measurements may not directly correspond to the outputs of the simulator and therefore require an extra layer of interpretation and analysis before the model predictions and the system observations can be compared. The observational error in Galform is of a particularly complex form, requiring considerable processing to transform all of the system observations to a comparable spatial and temporal resolution to the simulator outputs.

[5] **Initial condition and forcing function uncertainty** This corresponds to all of the other aspects of the simulator which need to be specified before the model may be evaluated. For example, the Galform simulator requires a full spatial specification of the arrangement of Dark Matter at all times in the development of the universe, and so we need to account for the uncertainty introduced as we do not know this configuration.

In this study, we will describe how we address each of these sources of uncertainty for the Galform project. We aim to be careful and thorough in each aspect of our quantification, but we must also recognise that, for a complex model such as Galform, uncertainty modelling is a process which is similar in many ways to the physical modelling process on which we are building. Quantification for each of the five sources of uncertainty that we must consider depends on complex scientific judgements over which different experts may have different views. Further, while there is much expert knowledge that is available and relevant for all aspects of these judgements, this information is held collectively over a wide community of experimenters, observationalists, theoreticians and modellers. Therefore, it is as misleading to talk of a definitive assessment of the uncertainty associated with Galform as it would be to talk of a definitive form for the Galform model itself. Assessment of uncertainty is an ongoing process for models which are, themselves, undergoing continuous development. Our account documents one iteration in this ongoing process, albeit one for which the uncertainty analysis is carried out to a much greater level of detail than is usual in this field (or indeed in most analyses of complex physical models in any area of application of which we are aware).

3.2 Linking the simulator with the system

We now introduce the general structure that we shall use to describe the relationship between the computer simulator and the physical system. We will describe this link in terms of the Galform simulator, but the ingredients are common to a wide variety of computer simulator analyses.

We denote by z the vector of observations that we shall use for this study¹. As we will describe in detail below, our choice for z will be the observed numbers of galaxies of various degrees of luminosity, assessed separately for younger and for older galaxies and expressed on the log scale. We describe the relationship between the observations, z , and the true physical system values, y , as

¹ z is not to be confused with Galaxy redshift.

$$z = y + \epsilon_{obs} \tag{1}$$

where ϵ_{obs} is the experimental error, which we take to be uncorrelated with y .

We are concerned as to whether the current theoretical understanding of Galaxy formation, as embodied in Galform, is consistent with observations z . Galform is represented as a function, which maps the inputs x to the outputs $f(x)$. The theoretical description involves the notion that when we evaluate Galform at the actual system properties, x^* say, then we should reproduce the actual system behaviour y . This does not mean that we would expect perfect agreement between $f(x^*)$ and y . Although Galform is a highly sophisticated simulator, it still offers a necessarily simplified account of the evolution of galaxies, and approximates the numerical solutions to the governing equations. The simplest way to view the difference between $f^* = f(x^*)$ and y is to express this as

$$y = f^* + \epsilon_{md}, \tag{2}$$

where we consider that ϵ_{md} is uncorrelated with f^* . Expressing our judgements about the likely size of the *model discrepancy*, ϵ_{md} , determines how close a fit between model output, f^* , and observation y we require for an acceptable level of consistency between theory and observation. Therefore, our judgements as to allowable model discrepancy are fundamental to the comparisons that we shall make.

We search for choices of input x for which the output $f(x)$ is sufficiently close to y that we would declare the observed output to be compatible with the predictions of the model, when we allow for model discrepancy. In practice, all that we can compare is $f(x)$ and z , which we do by combining (1) and (2). Achieving an acceptable match, for a particular input choice x , does not mean that the model is "correct" or that a choice of parameter values which achieve the match corresponds to the "true" value of the parameters, but simply that this version of the model will have met the challenge of reproducing an important observational aspect of the galaxy formation study within our agreed tolerance level. Similarly, identifying the whole collection of possible choices of inputs x which achieve an acceptable match is informative in identifying the ranges of parameter choices which are compatible with the given model and observations.

The form (2) is simple and intuitive, and is widely used in computer modelling studies. In our case, this corresponds to the natural approach in which we ask whether we could view Galform, with appropriate choice of inputs, as adequately reproducing the observed universe, within the tolerance set by the model discrepancy. In this account, we therefore ignore all of those additional aspects of our uncertainty modelling which would correspond to a more sophisticated analysis of model discrepancy, based, for example, on informed expert judgements as to the ways in which the Galform simulator is likely to evolve over the coming years. A detailed specification of such features would potentially be highly insightful, and might result in a much richer correlation structure across the elements of the discrepancy vector; see [10]. However, we have made the simplifying judgement that, as a first attempt to quantify uncertainties for Galform, it was better to focus on the most important large scale components of uncertainty. We shall describe in detail how we decompose structural

uncertainty into its leading ingredients. In the concluding sections, we will return to the issue as to how we might choose to build more structure into our description of discrepancy.

3.3 Bayes Linear Analysis

In this case study, we follow the *Bayes linear* approach to uncertainty quantification and analysis. This approach is relatively simple in terms of belief specification and analysis, as it is based only on mean, variance and covariance specifications which, following de Finetti, we take as primitive; see [8,9]. In this formulation, the probability of an event is the expectation of the corresponding indicator function. The appropriate updating rules for expectations and variances for a vector y , given a vector z are

$$\mathbf{E}_z[y] = \mathbf{E}(y) + \text{Cov}(y, z)\text{Var}(z)^{-1}(z - \mathbf{E}(z)), \quad (3)$$

$$\text{Var}_z[y] = \text{Var}(y) - \text{Cov}(y, z)\text{Var}(z)^{-1}\text{Cov}(z, y). \quad (4)$$

$\mathbf{E}_z[y]$ and $\text{Var}_z[y]$ are termed the *adjusted mean and variance of y given z* . Conditional expectation is the special case of belief adjustment where we base the adjustment on the indicator functions for a collection of events which constitute a partition. Bayes linear adjustment may be viewed as an approximation to a full Bayes analysis, or, more fundamentally, as the “appropriate” analysis given a partial specification based on whichever expectations we are both able and willing to specify. For a detailed treatment, see [11]. There are many areas of similarity between full Bayes and Bayes linear analyses. In particular, a full Gaussian specification for all of the relevant quantities would lead to similar updating formulae.

We have two basic reasons for choosing the Bayes linear approach for this study. Firstly, meaningful full prior probabilistic specification would be potentially very complex. For example, in relations (1) and (2), we have imposed the requirement that the two terms on the right hand side of each equation are uncorrelated. This already is a strong assertion, and we might well be reluctant to extend this to a judgement of full probabilistic independence between the corresponding terms. More generally, it may be reasonable to suppose that we can make expert judgements about the order of magnitude of the model discrepancy terms which are sufficient for us to make variance and covariance assessments across the various components. However, it seems unrealistic to imagine that we would be able to make the fine level probabilistic specifications over all combinations of model discrepancy outcomes which would be required for a full Bayesian analysis. Indeed, a more careful approach to specification might be to admit that even the mean and the variance of many of the quantities may only be specified with a degree of imprecision, and we will discuss the implications of such a view for our analysis in the appropriate sections below.

Our second reason for making this choice is purely technical. We wish to simplify the calculations that would be required for a fully probabilistic analysis. As we will show, the technical heart of our calculations is the iterative re-emulation of the Galform simulator within subspaces of the input parameters which are increasingly constrained by a series of complicated and highly non-linear boundaries. In order to render these calculations tractable, and to be able to visualise the results of our analysis, it is enormously helpful to exploit the simplifications of a Bayes linear analysis.

3.4 Emulation

We are interested in the behaviour of the Galform model over the whole of its specified input space. The substantial run time and the high dimensional input space combine to make direct exploration by model runs alone infeasible. We express our beliefs about the outputs of the model at locations in the input space that have not been previously evaluated by constructing an *emulator*. An emulator is a stochastic belief specification for a deterministic function [4–6, 12–14]. The emulator is much faster to evaluate than the simulator, so that we may explore the input space using the emulator, while taking into account the extra uncertainty that we have introduced by substituting emulator evaluations for simulator evaluations.

We construct our emulator for output i of the function $f(x)$ to have the form

$$f_i(x) = \sum_j \beta_{ij} g_{ij}(x) + u_i(x), \quad (5)$$

where $B = \{\beta_{ij}\}$ are unknown scalars, g_{ij} are known deterministic functions of x and $u(x)$, uncorrelated with B , is a weakly stationary stochastic process with constant variance. In our development, $g_{ij}(x)$ will usually be low order monomials in x , but they may take any forms that are appropriate to the problem at hand. The regression term on the right hand side of equation (5) expresses the global behaviour of the function, i.e. those aspects of the function about which we may learn by making a collection of function evaluations over a widely spaced, and roughly orthogonal design. We may exploit this global behaviour to make inferences about the value of the function across the whole input space. The process $u(x)$ represents localised deviations from this global behaviour near to x , and expresses those aspects of the behaviour of the function that we may only learn about by making function evaluations for which the inputs are close to x .

As we are employing the Bayes Linear approach, the emulator specification requires a mean vector and a variance matrix for B and values for the mean, variance and correlation function of u . A simple specification for $u(x)$ is to suppose, for each x , that $u_i(x)$ has zero mean with constant variance and where $\text{Corr}(u_i(x), u_i(x'))$ is a function of $\|x - x'\|$. The emulator is used to evaluate the expectation and variance of the function, for any input x and the covariance between the values of f at any pair of points x, x' . From (5), we extract these values as

$$\mu_i(x) = \text{E}(f_i(x)) = \sum_j \text{E}(\beta_{ij}) g_{ij}(x) + \text{E}(u_i(x)), \quad (6)$$

$$\begin{aligned} \kappa_i(x, x') &= \text{Cov}(f_i(x), f_i(x')) \\ &= \text{Cov}\left(\sum_j \beta_{ij} g_{ij}(x), \sum_j \beta_{ij} g_{ij}(x')\right) + \text{Cov}(u_i(x), u_i(x')). \end{aligned} \quad (7)$$

With high dimensional input spaces, it is common to find, for any output, f_i say, that a subset, $x_{[i]}$ say, of the inputs has the most influence in explaining the variation in the value of $f_i(x)$, where the subset $x_{[i]}$ may vary with i . In such cases, we may reform the emulator as

$$f_i(x) = \sum_j \beta_{ij} g_{ij}(x_{[i]}) + u_i(x_{[i]}) + w_i(x), \quad (8)$$

where $u_i(x_{[i]})$ has constant variance, and correlation function depending on $\|x_{[i]} - x'_{[i]}\|$, and $w_i(x)$ is a nugget term with constant variance over x , with $\text{Cov}(w(x), w(x')) = 0$ for $x \neq x'$. The collection $x_{[i]}$ is often called the *active variables* for f_i , and $w_i(x)$ is often referred to as a *nugget term* which expresses all of the variation in $f(x)$ which arises if we view the emulator $f(x)$ simply as a function of $x_{[i]}$.

There is some debate in the computer experiment literature as to whether it is preferable to put a lot of effort into constructing the regression terms in the emulator or whether it is better to construct a simple mean function and to place more weight on the residual process $u(x)$. Obviously, the best strategy is highly problem dependent. However, in this case and more generally, we prefer where possible to put as much detail as is feasible into the mean function. Our reasons for this preference are as follows.

[1] Many physical models, and Galform in particular, exhibit strong and physically interpretable monotonicities which are naturally expressed through the mean function.

[2] It is easier for the expert to assess whether the emulator formulation is consistent with informed scientific judgement about the behaviour of the function if a large proportion of the variability is expressed through regression terms.

[3] If much of the structure of the emulator is encoded in the regression function, then this simplifies various of the calculations that we need to make when comparing the model to observations and suggests very cheap approximations to calculations which would otherwise be very expensive if carried out using the full emulator across the whole of the input space.

[4] In our experience, the form of local process, $u(x)$, can be difficult to assess, even with large numbers of function evaluations. Partly, this is because there is a fundamental confounding between the location of the mean function, the size of the residual variance and the strength of the residual correlation. Partly, also, this is because any form of correlation function that we fit necessarily approximates the different degrees of smoothness of the function across different areas of the input space, and many methods of estimating smoothness parameters are potentially non-robust when applied to processes which do not fit exactly to the assumptions that are used to generate the fitting algorithms. Therefore, we prefer to model as much of the variation in the function as we can by the regression form, to reduce the residual variance as much as is feasible, and then to be fairly conservative in choosing the length of correlation that we shall impose. This has the effect of somewhat increasing our uncertainty away from the sampled input values, but, if the regression terms explain a sufficient proportion of the variation, then this does not have a large effect on our inferences.

In general computer experiments, we choose our form for the emulator by a combination of expert judgement based on physical intuition and experience with earlier versions of the model and, where appropriate, by preliminary experiments with fast approximate version of the simulator. In our case, we were able to make a collections of evaluations of the simulator, based on a Latin Hypercube design, which was sufficiently large to allow us to fit the emulator directly from our functional evaluations. Therefore we proceeded as follows, for each output that we chose to emulate.

Firstly, we carried out statistical model fitting, given the collection of runs, to select the deterministic functions g_{ij} , to assess the values of the coefficients B and to assess the

residual variance and covariance function, $u(x)$ and, where appropriate, to identify active subsets $x_{[i]}$. We then checked that the form of the emulator was physically meaningful. Finally, we carried out a diagnostic analysis on our emulator. This has two aspects. Firstly, we inspected the estimated residuals at the locations at which we had evaluated the function. Secondly, we carried out a new diagnostic design, where we made a collection of additional function evaluations and checked that our emulator gave an acceptable representation of uncertainty as judged by the new evaluations, looking, in particular, for new evaluations which fell at a large standardised distance from their expected values. We will give details of each of these stages in the construction of our emulators below.

3.5 History Matching

The aim of this study is to identify the set of input values \mathcal{X}^* for which the evaluation of $f(x)$ is sufficiently close to the observations z that we may consider the observed values for the output quantities to be compatible with the predictions of the model, when we allow for model discrepancy and observational error. As we are unable to evaluate $f(x)$ over the whole input space, we will need to estimate the set \mathcal{X}^* , by identifying all values of x for which there is good reason to suppose that we would obtain an acceptable match were we to make such an evaluation of $f(x)$, along with obtaining a substantial collection of actual evaluations of the function which actually do yield acceptable matches and which may then be used to explore the match between other aspects of the Galform output and the corresponding observational information.

We refer to the process of identifying the collection \mathcal{X}^* as *history matching*. This terminology is common in various applications, and in particular in oil reservoir modelling, where it refers to the process of adjusting the inputs to a simulator of an oil reservoir until the output closely reproduces features such as the historical oil production and pressure profiles at all of the wells. The emphasis on identifying all of the possible matches to observation is ours. While it is always of interest to know how wide a range of input choices may give rise to history similar to that which has been observed, it may be difficult to find any history matches, for a complex and slow to evaluate simulator, never mind the collection of all such matches. Therefore, pragmatically, reservoir engineers often stop when a few matches, or even just one, have been obtained.

History matching may be compared to the more familiar problem of model *calibration* in which we suppose there is a single "true but unknown" value x^* and our objective is to make probabilistic statements as to this value, based on a prior specification for x^* , the collection of model evaluations and the observed history. While calibration and history matching are thematically related, we can see that they are fundamentally different. For example, calibration will always result in a proper posterior distribution over the input space, while history matching might lead to the conclusion that the collection of acceptable matches was empty. It would be of great interest to find that the set \mathcal{X}^* was empty in the Galform study, as that might suggest possible defects in the general theory underlying the simulation process. However, in this study, we do find a collection of good fits to the observations.

Our general view is that history matching is always of interest for assessing computer

models and calibration sometimes is. Even when we wish to carry out a model calibration, we consider that it is often good practice first to carry out a history match, partly to see whether such a match is achievable, and partly to reduce the size of the input space over which the calibration exercise will need to be performed. In our case, we do not follow history matching with calibration, although we do suggest informal measures for selecting "likely" choices for x^* if we subsequently decide that such an assessment is meaningful.

Our approach to history matching is based on the assessment of certain *implausibility measures* as we now describe. An implausibility measure is a function defined over the input space which, when large, suggests that the match between model and system would exceed our stated tolerance. We may build this up as follows, for a single output $f_i(x)$. For a given choice, x^* , we would like to assess whether the output $f_i(x^*)$ differs from the system value y_i by more than the tolerance that we allow in terms of model discrepancy. Therefore, we would assess the standardised distance

$$\frac{(y_i - f_i(x^*))^2}{\text{Var}(\epsilon_{md:i})}$$

In practice, we cannot observe y_i and so we must compare $f_i(x^*)$ with the observation z , introducing measurement error, with corresponding standardised distance

$$\frac{(z_i - f_i(x^*))^2}{\text{Var}(\epsilon_{md:i}) + \text{Var}(\epsilon_{obs:i})} \quad (9)$$

However, for most values of x , we are not able to evaluate $f(x)$ so we use the emulator and compare z_i with $\text{E}(f_i(x))$. Therefore, the implausibility function is defined as

$$I_{(i)}^2(x) = \frac{(\text{E}(f_i(x)) - z_i)^2}{\text{Var}(\text{E}(f_i(x)) - z_i)} = \frac{(\text{E}(f_i(x)) - z_i)^2}{\text{Var}(f_i(x)) + \text{Var}(\epsilon_{md:i}) + \text{Var}(\epsilon_{obs:i})} \quad (10)$$

When $I_{(i)}(x)$ is large, this suggests that, even given all the uncertainties present in the problem, we would be unlikely to view as acceptable the match between model output and observed data were we to run the model at input x . Therefore, we consider that choices of x for which $I_{(i)}(x)$ is large can be discarded as potential members of the set \mathcal{X}^* . We discard regions of the input space by imposing suitable cutoffs on the implausibility function. We will discuss appropriate choices below.

In our comparisons, we have a separate implausibility function for each output that we use for history matching. We may either choose to make some intuitive combination of the individual implausibility functions as a basis of eliminating portions of the input space, or we may construct the natural multivariate analogue of the scalar function, of form

$$(z - \text{E}(f(x)))^T (\text{Var}(z - \text{E}(f(x))))^{-1} (z - \text{E}(f(x)))$$

The multivariate form is more effective for screening the input space, but it does require careful consideration of the covariance structure for the various ingredients of the uncertainty specification. In our account, we will use both univariate and multivariate implausibility measures as appropriate.

History matching is intended to be an iterative process. We begin by emulating Galform over the whole input space. We evaluate our implausibility measures over the whole space and remove from the space all input choices for which the implausibility measure is large. We then re-sample within the remaining input space and re-emulate Galform within the reduced space. This is termed *refocusing*. We then recalculate the implausibility measures over the reduced space and again remove those parts of the subspace for which the new implausibility measure is large. We re-sample within the further reduced space, re-emulate and again re-assess the implausibility measures, further reduce the input space and continue in this fashion until we run out of time, budget or ability to further reduce the input space, at which time we look to generate a large number of acceptable runs from the remaining space. The reasons that we may hope to further reduce the acceptable space at each iteration are firstly that we produce a higher relative density of runs at each stage, so that emulation is more effective, secondly that we may expect the function to become smoother and so easier to emulate as we reduce the area of the input space, and thirdly because, when we have accounted for much of the uncertainty related to the most important active variables, then variables which did not account for much of the variability in the original emulation may take on larger importance and therefore allow us to resolve more of the uncertainty of the function.

In this study, we refocused four times, and then carried out a fifth set of evaluations which produced a large number of runs which gave good matches to observations. One of the most interesting features of this study, from our viewpoint, is that we had never carried out history matching to this level of detail before, usually being content to stop after a single refocusing stage, and it was very instructive to see the progressive reduction of the input space from the repeated refocusing. This continued refocusing is very useful, but it also brings its own complications, as the only way in which we can determine whether an input value lies within our retained collection of potential history matches is by applying each implausibility function in turn and seeing whether each such evaluation is small enough for the input choice to be retained. This raises practical computational issues, which makes it important to have fast approximate methods to screen the input space, and also raises basic questions about practical visualisation methods to help us to represent and interpret the shape of the input space which we have retained.

4 The Galform Model

We have described the physical motivation and the general structure of the Galform model in section 2. Here we identify the specific attributes of Galform that are relevant in a Computer Model Uncertainty Analysis, and link these to the formal definitions discussed in section 3. We discuss features of the Dark Matter simulation, and give a detailed description of the Galform model itself, including the inputs and outputs used in this analysis.

4.1 The Dark Matter Simulation

Before the Galform Model can be run, a large Dark Matter simulation must be performed, known as the Millennium simulation [19]. This simulation only models the behaviour of mass acting under the gravitational force, from the beginning of the Universe until the current day. As described in section 2, small quantum fluctuations in the mass density of the early Universe grow, collide and eventually form a collection of immense lumps of mass, referred to as Dark Matter Haloes. It is the detailed history of the growth and mergers of these haloes that is a required input for the Galform Model.

The Millennium simulation models approximately 10^{10} massive particles under the influence of gravity, covering a volume corresponding to $(1.63 \text{ billion light years})^3$ at the present day. It is, understandably, extremely computationally expensive and one evaluation takes about 3 months using a supercomputer (see the Virgo consortium for details). Although the Millennium simulation requires the choice of 5 cosmological parameters that it would be in principle interesting to investigate, the substantial run time of the model resulted in only one evaluation being performed, with the cosmological parameters set at values consistent with recent measurements from the WMAP satellite [18]. The possibility of future runs with different parameter settings would allow assessment of the consequences of uncertain cosmological parameters; however, this will not be available for the current study.

4.2 The Galform Model

Galform is the model of interest in this Case Study. It takes the results of the Dark Matter simulation, specifically the merger histories of the Dark Matter Haloes, and then models the far more complicated behaviour of baryonic (i.e. normal) matter. It is the baryonic matter that is responsible for the more intricate processes involved in galaxy formation such as star formation, gas accretion, and the feedback from black holes and supernovae.

As the Millennium simulation covers a substantial volume, its results are split into 512 sub-volumes, each of which are simulated using the Galform model with edge effects dealt with using periodic boundary conditions. This splitting of the total volume was performed to increase computational efficiency as it allows simple parallelization of the Galform model across multiple processors. The run time for one evaluation of the Galform model on a single sub-volume is approximately 30 minutes. After discussions to initiate the collaboration, the Galform group provided shared access to a cluster of 256 processors (composed of 128 dual processor Sunfire V210s, each processor being an UltraSparc IIIi with a clockspeed of 1 GHz and with 1 GByte of RAM per processor). Previous attempts by the cosmologists to calibrate Galform focussed on the first 40 sub-volumes out of 512, and we follow this approach here while taking account of the uncertainty this generates. Examining the differences between Galform output from different sub-volumes allows an assessment of this uncertainty as is described in section 6.1.2.

We treat Galform as a computer model in order to learn about which inputs x will give rise to acceptable matches between the outputs $f(x)$ and the observational data z . If such inputs can be found, and the Galform model is shown to give a good description of the

properties of galaxies in our observed Universe, then this would provide evidence for both the existence of Dark Matter and for the theory of hierarchical structure formation. If, conversely, no acceptable inputs can be found then this would cast serious doubt on the validity of these popular theories.

The emulation and History Matching process that we employ will also help to improve the cosmologists’ understanding of the model. Emulators give useful insights into the structure of the Galform function $f(x)$, while History Matching identifies regions of input parameter space that give comparable matches to the output data. The insights gained from use of these techniques will be vital both when considering coding improvements to the current model, and when performing basic diagnostic checks on the physics of the model itself.

4.3 Galform: Physical Details

We now outline some relevant technical details of the GALFORM code. An extended description and discussion of the implementation can be found in [1] “A galaxy formation primer”. Here we provide a brief summary. In essence, the model consists of a set of modules describing the distinct physical processes. Each has its own associated parameters, and we discuss each of the components in turn.

1. **Dark matter merger trees.** These are extracted from the “Millennium” dark matter simulation [19]. This is a full numerical simulation of the growth of dark matter structures in the universe from cosmological initial conditions. The initial spectrum of density fluctuations is set to be consistent with the WMAP satellite observations of the cosmic microwave background [18]. The subsequent evolution involves solving the gravitational N-body problem for a collection of 10^{10} particles. The computations took several months on state of the art super computers at the Max Planck Society’s Rechenzentrum in Munich, Germany. Fortunately, this part of the model need only be solved once, and the main part of the GALFORM code can then be applied to populate the dark matter haloes with galaxies. This approach improves accuracy over previous analytic approximations to gravitational structure growth, but means that we must fix the cosmological parameters for our model. In future, improved analytic modelling of the merger trees will allow us to include the uncertainty in the cosmological parameters. For now, cosmological parameters are fixed to the canonical year 3 observations of WMAP in which $\Omega_b = 0.045$, $\Omega_M = 0.25$, $\Lambda = 0.75$ and $\sigma_8 = 0.9$ at the present day. The model assumes $H_0 = 0.73$, although we quote luminosities and space densities in term of $h = H_0/100\text{kms}^{-1}$ so that this dependence is explicit.
2. **Gas Accretion and Cooling.** As dark matter haloes grow, the gas that they contain cools and flows to the centre. This occurs at different rates depending on the mass of the halo, and the rate at which the halo mass grows. The supply of gas is determined by computing the mass of gas for which the cooling timescale is less than the halo, and the mass of gas which has had sufficient time to cool and fall to the centre. The gas is supplied to the central galaxy as the smaller of these two quantities. Further explanation of this part of the process is given in [1, 3].

However, the version of the code used in Bower et al 2006 (B06) [2] made several important changes to previous versions of the GALFORM code. One of these was to emphasise the distinction between haloes for which the gas supply is limited by the rate of cooling (henceforth “hydrostatic” haloes) and those haloes for which the free-fall timescale is the limiting factor (henceforth “rapid cooling” haloes). In the B06 model, it is assumed that energy from the central black hole can only offset the cooling in hydrostatic haloes. Because this process is poorly understood, we introduce a parameter α_{cool} that determines the exact ratio of timescales at which this distinction is made.

3. **Star Formation.** As the hot gas cools or is accreted by a halo, it falls towards the galaxy at the centre building up a reservoir of cold gas. This gas provides the fuel for the formation of further stars. The code assumes that the star formation rate is related to the dynamical timescale of the galaxy, and its mass of gas, giving

$$\dot{m}_* = \epsilon_* \left(\frac{m_{\text{cold}}}{\tau_{\text{disk}}} \right) \left(\frac{v_{\text{disk}}}{200 \text{kms}^{-1}} \right)^{\alpha_*}$$

where \dot{m}_* is the star formation rate, m_{cold} is the mass of cold gas, τ_{disk} is the disk dynamical time and v_{disk} is the disk rotation speed. α_* and ϵ_* are parameters that control the rate of star formation and its dependence on galaxy mass.

In B06, an additional mode of star formation is also considered. If the disk becomes too massive, it becomes susceptible to warps that grow, funnelling gas to the center of the galaxy. Such secular evolution may generate many of the bulges that are observed. In the model it is assumed that instabilities occur if the disk’s gravity exceeds the stabilising gravity of the halo. The threshold at which this occurs is set by the parameter f_{stab} , at which point the disk stars are added to the galaxy’s bulge and the disk gas is consumed in a burst of star formation.

4. **Feedback - from supernovae.** Soon after the most massive stars form, they explode in powerful supernova explosions. These are thought to be responsible for preventing the efficient formation of stars in small galaxies - as the stars form, gas is driven out of the system by the supernovae. We model feedback from supernovae by assuming that the ratio of material expelled from the galaxy into the halo to that formed into stars is given by the ratio β , where

$$\beta = (v_{\text{disk}}/v_{\text{hot}})^{-\alpha_{\text{hot}}} \tag{11}$$

where v_{hot} and α_{hot} are poorly constrained parameters. We allow v_{hot} to take different values for quiescent and burst star formation which we denote as $V_{\text{hot,burst}}$ and $V_{\text{hot,disk}}$.

The gas that is driven out of galaxies flows into the halo, but does not immediately become available for cooling. The timescale on which the gas becomes available is determined by the parameter α_{reheat} . If this is unity, and cooling is efficient, ejected gas will be allowed to fall back into the galaxy on the dynamical timescale.

5. **Galaxy mergers.** When dark haloes collide, the galaxies at their centers do not immediately merge. Rather their relative motion slowly decays due to dynamical friction. This process is discussed extensively in [3]. The merging time is set by an overall normalisation parameter $\tau_{0,\text{mrg}}$.

If the time since the halo was accreted is less than the merging time, the galaxy from the “satellite” galaxy orbits inside the larger one. Such satellite galaxies do not collect any gas from the halo, and so star formation quickly subsides as the cold gas reservoir is exhausted. If the time since accretion exceeds the merging timescale, the galaxy merges with the central galaxy in the parent halo. If the mass ratio of the galaxies exceeds f_{ellip} , this can cause disturbance to the underlying galaxy, transforming it from a spiral type galaxy to an elliptical one. This morphological transformation may be associated with a burst of star formation. If the mass ratio exceeds f_{burst} , there is no morphological transformation, but a burst of star formation still occurs.

6. **Black holes and their feedback.** The model assumes that black holes grow through three distinct channels: (i) by black hole - black hole mergers when the parent galaxies merge; (ii) by accretion of gas that is funnelled to the galaxy center during bursts of star formation (these being driven either by mergers or disk instabilities); (iii) by diffuse gas accretion from hydrostatic haloes (i.e., as a result of “radio mode” feedback).

The star burst driven accretion results in luminous quasars, but the current model assumes that these events do not contribute to the feedback. The parameter F_{bh} controls the amount of gas that is accreted by the black hole in these events. The feedback from “radio mode” accretion is, however, of key importance. The mass growth of the black hole is determined from the energy output required to counter-balance cooling of the halo, i.e. we implicitly assume that the mass accretion rate increases until the net cooling rate decreases to zero. This is an important caveat. Accretion onto black haloes, although an abundant source of energy has limits. We therefore limit the maximum energy output to be less than $\epsilon_{\text{Edd}}L_{\text{Edd}}$ where L_{Edd} is the Eddington luminosity of the black hole and ϵ_{Edd} is an adjustable parameter. Current models for black hole accretion suggest that ϵ_{Edd} is of order 1%.

7. **Reionisation** At very early times, the majority of gas in the universe is neutral (and the universe is opaque to ultra-violet light). As stars and quasars form in abundance, the universe quickly ionizes. This creates an additional form of heating that may be extremely important in very low-mass galaxies. The details of this process are very important for understanding the paucity of dwarf galaxies that orbit in the milky-way halo. However, we are here concentrating on the properties of much more massive systems where these effects are less significant and it is sufficient to parameterise this process by two parameters, z_{cut} and v_{cut} . Here, z_{cut} defines the redshift at which reionisation occurs: at lower redshifts, gas cooling is prevented in haloes with circular velocity below v_{cut} .

Input Parameters	symbol	min	max	Initial Vars	Active W1	Process Modeled
vhotdisk	$V_{\text{hot,disk}}$	100	550	x	x	SNe feedback
vhotburst	$V_{\text{hot,burst}}$	100	550	x	x	
alphahot	α_{hot}	2	3.7		x	
alphareheat	α_{reheat}	0.2	1.2	x	x	AGN feedback
alphacool	α_{cool}	0.2	1.2	x	x	
epsilonSMBHEdd	ϵ_{Edd}	0.004	0.05			
epsilonStar	ϵ_{\star}	10	1000	x	x	Star Formation
alphastar	α_{\star}	-3.2	-0.3			Disk stability
yield	p_{yield}	0.02	0.05		x	
tdisk	t_{disk}	0	1			
stabledisk	f_{stab}	0.65	0.95	x	x	Galaxy Mergers
tau0mrg	$\tau_{0,\text{mrg}}$	0.8	2.7			Reionisation
fellip	f_{ellip}	0.1	0.35			
fburst	f_{burst}	0.01	0.15			
FSMBH	F_{bh}	0.001	0.01			Reionisation
VCUT	v_{cut}	20	50			
ZCUT	z_{cut}	6	9			

Table 1: Table of Parameter Ranges (these ranges were converted to -1 to 1 for the analysis), including the initial variables considered, those analysed in Wave 1 and the physical processes that each parameter relates to.

4.4 Inputs

The Galform model has a total of 17 inputs that relate to various uncertain physical processes involved in galaxy formation which were described in section 4.3. All 17 inputs along with their considered ranges are shown in table 1. Also shown are the variables that are deemed as Active in Wave 1 of our analysis: this will be discussed in more detail in section 5.1.4.

To make one evaluation of the Galform model, single values for each of the 17 inputs must be chosen. We write this vector of 17 inputs as x . The fundamental question is what values of x will give acceptable agreement between model output $f(x)$ and observational data z .

4.5 Outputs

Galform provides several different sets of output data related to various physical characteristics of the simulated galaxies. Observational data of differing degrees of accuracy are available for comparison with the Galform model output, the most important of these being the bj and K Luminosity Functions. These Luminosity functions give the number of galaxies of a certain luminosity, per unit volume, as a function of luminosity, with the bj function rep-

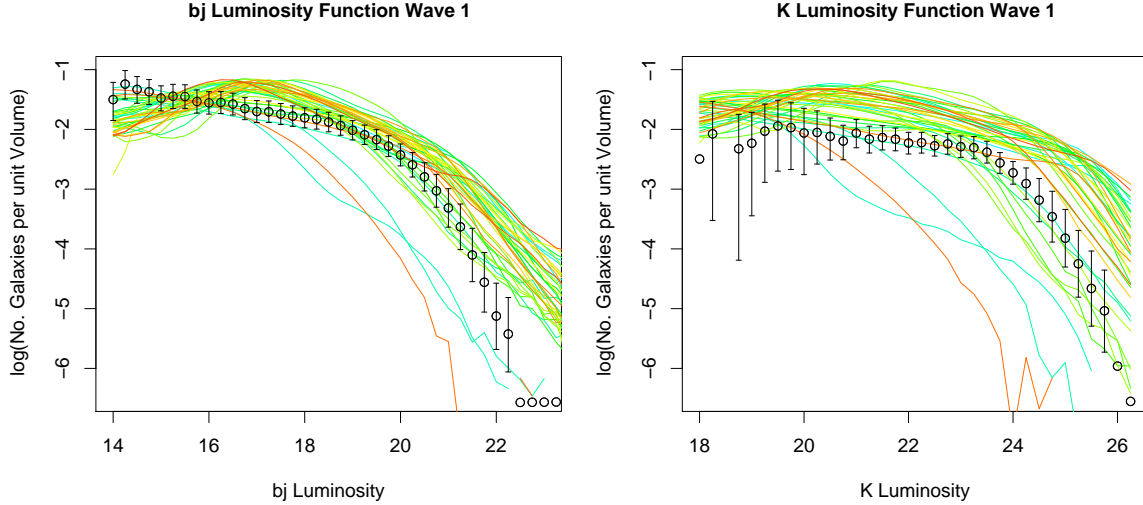


Figure 1: The observed bj (left) and K (right) Luminosity Functions giving the number of galaxies of certain luminosity, per unit volume. The data are shown as the black points, along with 2 sigma intervals representing all relevant uncertainties identified in section 6. Also shown are the outputs from 40 runs of the Galform model (the coloured lines).

representing bluer (mainly younger) galaxies and the K function redder (mainly older) galaxies. The ‘bj’ and ‘K’ are purely labels identifying the wavelength or colour of the light measured (blue and infrared respectively).

Figure 1 shows the bj and K luminosity function data (black dots) along with all relevant uncertainties discussed in section 6, on a \log_{10} scale. The y-axis gives the log of the number of galaxies per unit volume, while the x-axis represents the luminosity with brighter galaxies at higher values. Figure 1 also shows the outputs of the first 40 runs of the Galform model (the coloured lines), demonstrating that performing runs at random is unlikely to yield any useful matches to the data.

Note the apparent ‘break’ in the luminosity functions at luminosity of approximately 19.5 for the bj graph and 23.5 for the K graph: to the left of this point the function is approximately linear, but after this point there is a sharp decrease in the numbers of galaxies observed. This paucity of bright galaxies has been the subject of much interest within the cosmology community as it has been a difficult feature to reproduce (and is one of the reasons why the observations are most commonly viewed on a log scale). The current version of Galform was the first model to include the physics involved in Active Galactic Nuclei, and hence was able to model this break in the luminosity function (see section 2 for more details).

The Luminosity Function data set represents the most accurately measured observational data available and is seen as the benchmark by which models of galaxy formation are judged. Even if a particular galaxy formation model performs well with respect to other data sets, if it does not match the Luminosity function to an acceptable level then that model will be discarded.

For these reasons, it was decided to focus our analysis on identifying the regions of input space that give rise to matches between the model output and the b_j and K observed luminosity functions. Additional data sets could then be used at a later date to restrict the input space further.

5 First Wave Analysis

5.1 The Wave 1 Emulator

5.1.1 General Designs for Computer Model Experiments

We want to explore the input space of Galform in order to understand which choices of input configuration will give rise to acceptable matches between the outputs of the model and the observed data. We have to explore a high-dimensional input space of a model which takes a significant amount of time to run. Therefore the design for the set of input configurations where evaluations of the model will be performed is very important: this is a general problem that arises in most Computer Model analyses [7, 16, 17]. An important attribute of such a design is that it should be space-filling: we want to ensure that no two runs are close together, and hence to maximize the coverage of the input space. Another desirable property is that the design is at least approximately orthogonal (where possible), as we will be fitting various polynomials to the outputs when constructing the emulator. Various designs have been discussed in the Computer Model literature [17], with a popular choice being the Maximin Latin hypercube design. An n point Latin Hypercube design is constructed by dividing the range of each of the input variables into n equal intervals. Points are placed so that one point will occupy each of the n intervals, for each input variable. Maximin Latin Hypercube designs are constructed by generating many Latin Hypercube designs and selecting the one that has the maximum ‘minimum distance’ between points. They are approximately orthogonal designs and suffer no projection issues as any lower dimensional projection remains a Latin Hypercube. They are therefore of use for Computer Model experiments such as Galform, where large batches of runs are to be evaluated, and we expect to fit the emulator within appropriate subspaces of the full input space.

5.1.2 The Wave 1 Design

The first stage in the collaboration concerned History Matching using a smaller number of input variables than were present in the full Galform model, in order to demonstrate the methodology in a simplified version of the problem. As the collaboration progressed we extended our aims to include an analysis of the full model with all 17 input parameters. This evolution in priorities has had an impact on the general structure of the analysis, as will be noticeable from the initial design choices described here, as well as in many later areas.

When considering the initial design, expert judgements were used to identify a subset of the inputs which would have either significant effects on the b_j and K luminosity function

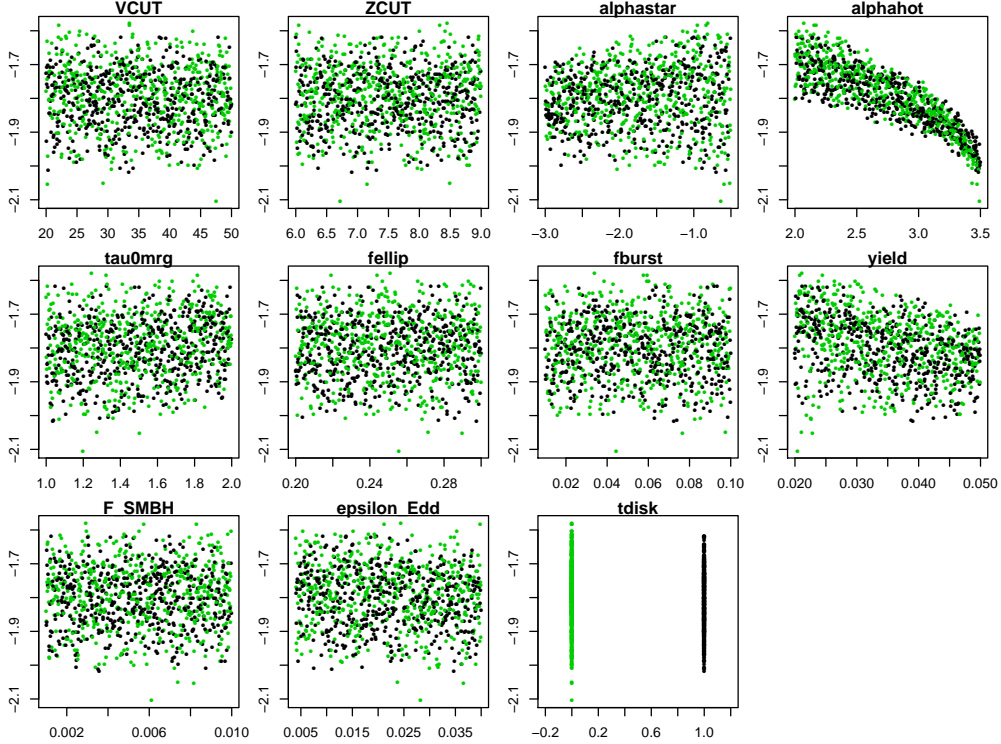


Figure 2: Main effects plots for the bj output at luminosity = 18.75 against each of the 11 inputs thought to be inactive. The input $alphahot$ is clearly responsible for a large part of the variation and was hence promoted to the active set. Note that $yield$ was also promoted at the Cosmologist’s request, and in Wave 4 both $alphastar$ and $tau0mrg$ are found to be active. The green and black points correspond to the discrete parameter $tdisk$ being set to “off” or “on” respectively.

outputs that we are considering, or be of physical interest to the cosmologists (unless otherwise stated, all expert judgements in this study were made by Richard Bower). This set of 6 of the inputs to be used in the first analysis are shown in the ‘Initial Variables’ column of table 1. The remaining 11 variables were thought to be either of less physical interest, or to be not directly related to the physical processes relevant to the luminosity functions. Another important issue that had an effect on the initial designs was that, when the Galform project began, it was impossible to run the model while varying more than 11 input parameters simultaneously due to technical issues with the code. Therefore, we constructed two maximin Latin Hypercube designs: the first was over the 6 inputs identified as important and was used for initial analysis, and subsequently the second design was evaluated over the 11 inputs thought to be less significant.

Consideration of the two sets of runs provided useful insights into features of the model that would be used when performing the full analysis over all 17 inputs. An initial analysis of the first set of runs (which we will not report on here), suggested that acceptable matches

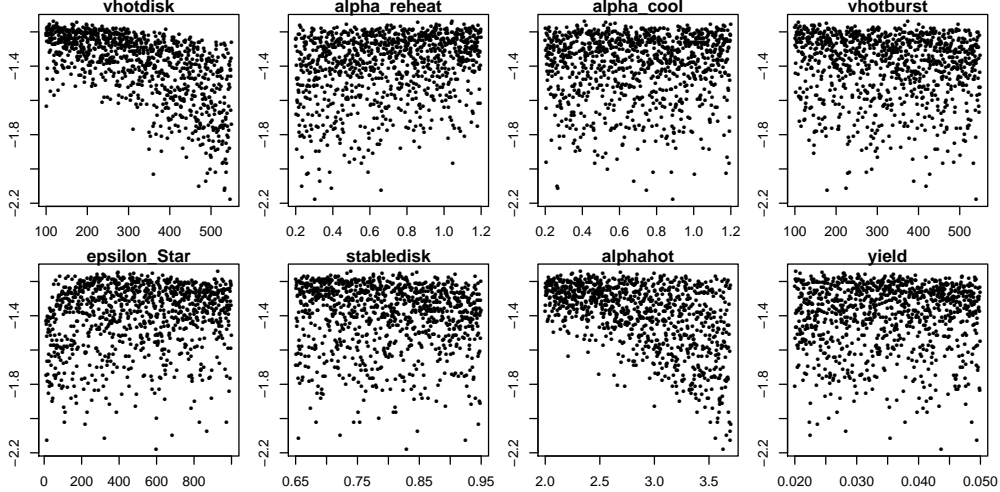


Figure 3: Main effects plots of b_j output at luminosity = 17, for the 993 completed runs over the 8 possibly active inputs. Note the clear effect of inputs vhotdisk and alphahot: these along with epsilonStar, alphareheat and vhotburst were eventually chosen as the active variables for this output (see section 5.1.4).

could, most likely, only be found for extremely low values of the 5th input parameter epsilonStar, with the Galform function decreasing rapidly at such values. This made intuitive sense as the relevant physical process is dependent upon the inverse of epsilonStar (see section 4.3). We therefore reparameterised this input as epsilonStar^{-1} for all subsequent analysis.

Comparison of the variance of the outputs in each data set implied that one parameter (alphahot) out of the 11 initially discarded inputs, had a clearly significant effect on the luminosity functions, and after careful consultation, this input was promoted into the active group. At this point, the cosmologists requested that the parameter “yield” also be promoted, as recent physical evidence had suggested that the value assigned to this parameter in previous analyses (0.02) was too low, and hence the cosmologist were interested in finding acceptable matches with a higher yield value.

Figure 2 shows the main effects plots for the b_j output at luminosity = 18.75, for the 1000 runs over the 11 inputs initially thought to be of less interest. It clearly shows the impact of varying alphahot and to a lesser extent, yield, both of which were promoted for the full analysis.

Once these initial investigations were complete, we were ready to proceed with the analysis of the full Galform model. We constructed two large Latin Hypercube designs: the first over the 8 possibly active variables (shown in table 1), and the second over the 9 inactive variables. The number of points in each design was chosen by consideration of run time, computational resources and the probability of runs crashing. As we had decided to analyse the first 40 sub-volumes of the model (in order to compare with previous work performed by the cosmologists), and had shared access to a cluster of 256 processors, we decided to

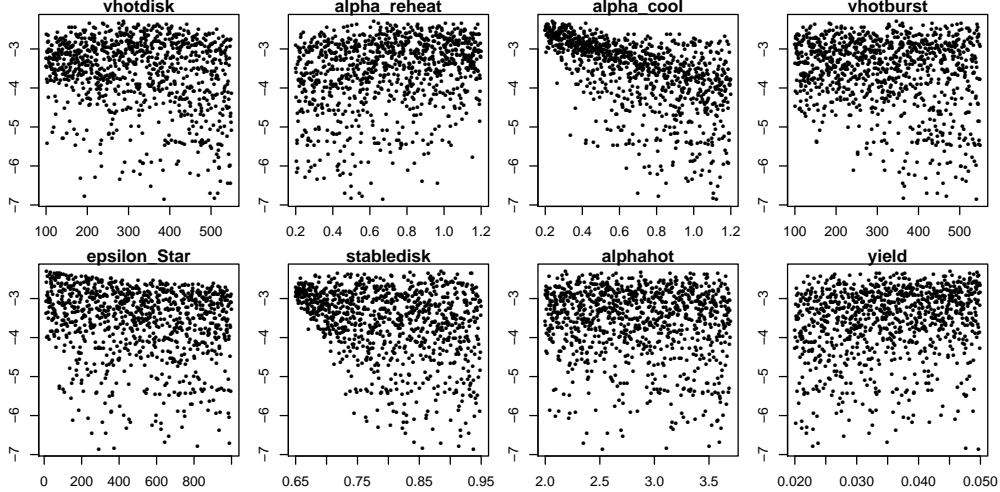


Figure 4: Main effects plots of K output at luminosity = 25.75, for the 993 completed runs over the 8 possibly active inputs. Note the clear effect of inputs alphacool, vhotburst and stabledisk: these along with vhotdisk and alphareheat were eventually chosen as the active variables for this output (see section 5.1.4).

run both designs with 1000 points each. The first of these was used to construct the Wave 1 emulator (see section 5.1.3), and the second was required to assess the uncertainty due to the set of 9 inactive parameters (see section 6.1.1). Unfortunately, multiple runs crashed and had to be repeated, often more than once. These crashes were caused by technical reasons concerning the cluster, and not related to physical attributes of the runs in question. Due to time constraints only 993 of the first batch of runs were completed, while all 1000 of the second batch finished successfully. For illustration, Figures 3 and 4 show main effects plots for the b_j and K outputs at luminosity 17 and 25.75 respectively, for the first batch of 993 runs against the 8 possibly active input parameters.

We are performing a History Match for Galform. For such a match we do not need to analyse every output of the model. At each stage, it is sufficient to remove parts of the parameter space if the outputs fail to match a carefully chosen subset of the observations. At the final stage, we will need to check that our acceptable matches are also in adequate agreement with those features of the output which haven't been used to achieve the history match. Therefore, we choose a subset of the outputs that are both straightforward to emulate at a sufficient accuracy, and are also informative regarding the inputs in that they can be used to discard large regions of the input space. 7 such outputs were chosen, that captured the main features of the luminosity function, and these are shown in figure 5 along with the full b_j and K luminosity outputs from the first batch of 993 runs over the 8 active parameters. In later waves of the analysis, more outputs were used.

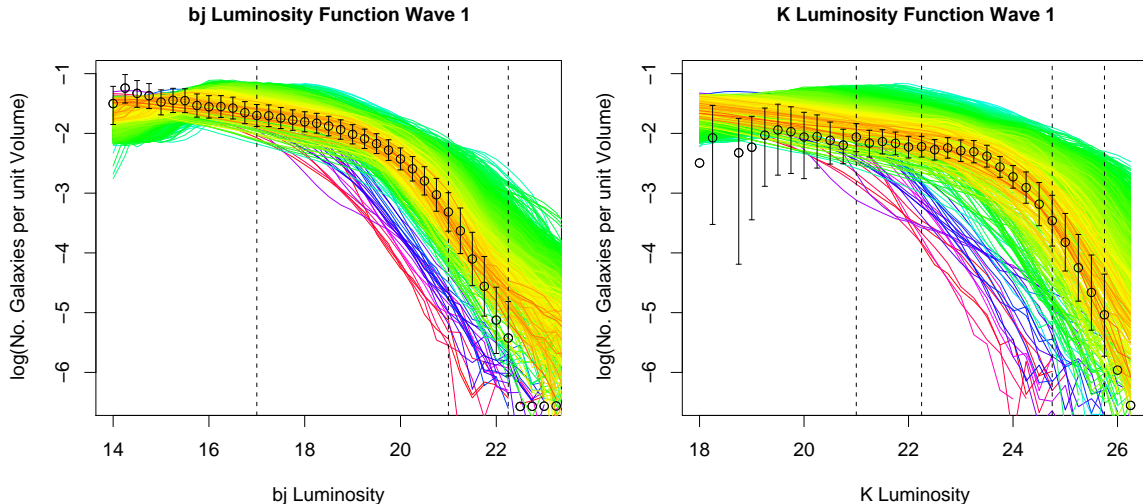


Figure 5: The bj and K luminosity outputs from the 993 Wave 1 runs of the model (coloured lines). The vertical lines show the 7 outputs chosen for emulation, and the black points show the observational data. The error bars show 2 sigma intervals and incorporate all sources of uncertainty including model discrepancy and the observational errors as are described in sections 6.1 and 6.2

5.1.3 Emulator Construction

We now describe the techniques that we used in the construction of a univariate emulator, elaborating on the discussion in section 3.4. An emulator is a stochastic belief specification for a deterministic function. It encapsulates our beliefs about the output of the function $f(x)$ for any input point x . As we are using a Bayes Linear approach, the emulator will give, for each input parameter setting x , an expectation and variance of the function: $E(f_i(x))$ and $\text{Var}(f_i(x))$. This uncertainty will be small at points close to evaluated model runs, and larger at points far from known runs. Our emulator will be constructed using a large, space filling set of evaluations of the model, as described in section 5.1.2.

Often in Computer Model experiments several of the input parameters x_i have little effect on the output of interest. This can be incorporated into the emulation process in the following way. Following the discussion given in section 3.4, we define $x_{[A_i]}$ to be the set of Active Variables: those that have strong effects on the outputs that are to be emulated, and model their effects on $f_i(x)$ in detail. The remaining inputs (referred to as inactive) are treated as contributing a noise term to the emulator. Hence the emulator for component i of $f(x)$ would now have the form:

$$f_i(x) = \sum_j \beta_{ij} g_{ij}(x_{[A_i]}) + u_i(x_{[A_i]}) + w_i(x). \quad (12)$$

Here the functions $g_{ij}(x_{[A_i]})$ and the weakly stationary process $u_i(x_{[A_i]})$ depend only on the active variables (which may be a much smaller set than the full set of input parameters). The

effects of the inactive parameters are described by the $w_i(x)$ term, referred to as a nugget, which is modeled as white noise and hence is uncorrelated with any other terms in the emulator (and with itself at different input parameters settings). The use of active variables can greatly simplify the emulation process and subsequent analysis by reducing the dimension of the input space and hence of the computations that we must make, with usually only a small impact on the accuracy of the results. Choosing the set of active variables is done by a combination of expert elicitation and data analysis on the ensemble of runs. As we have a large collection of runs, we construct our emulator using data analytic techniques, checked against physical intuition. While all such data analytic methods involve a certain amount of trial and error based on features such as careful residual analysis, the general structure of our approach is as follows. Active variables were selected using backwards stepwise elimination, starting with a model containing all the linear terms in x . Once the active variables were specified, the functions $g_{ij}(x_{[A_i]})$ were determined. Although expert judgements may suggest appropriate forms for these functions, a simple choice is to use monomials of low order, less than or equal to k . The order k will depend on the number of active variables and on the number of runs, which determines the level of detail at which we may describe the function. To identify which set of monomials will be kept, backwards stepwise elimination, starting with the full polynomial over all active inputs $x_{[A_i]}$ of order k , can be used to remove less significant terms from the model. Once the final form of the model is determined for each output $f_i(x)$, the coefficients $B = \{\beta_{ij}\}$ can be assessed using Ordinary Least Squares (assuming uncorrelated errors), as we have a sufficiently large collection of model evaluations that such data analytic techniques will result in small variances on the regression coefficients and generally acceptable results from OLS fitting. Therefore, we would expect such results to overwhelm prior judgements. However, any substantial contradictions between the data and the qualitative form of such judgements always requires further investigation.

All that remains is to specify the precise form of the two contributions $u_i(x_{[A_i]})$ and $w_i(x)$ to the residual process. It is often reasonable to assume a constant variance such that $\text{Cov}(u_i(x), u_i(x')) = \sigma_{u_i}^2 r(\|x - x'\|)$, where $r(\|x - x'\|)$ is an auto-correlation function. Various forms for $r(\|x - x'\|)$ are available, such as the Gaussian or Matern functions. Any sensible choice of $r(\|x - x'\|)$ will involve certain parameters related to the width and shape of the correlation function, and estimation of these parameters can be a difficult task. However, it should be emphasized that these parameters are representations of our subjective assessment of the smoothness of the function and an accurate assessment of them is not necessarily meaningful, and nor is it required in order to construct an emulator of sufficient accuracy for our needs. The variance of $u_i(x)$ and the variance of the nugget $w_i(x)$ can be assessed in a variety of ways, but are usually linked to the estimate for the residual variance obtained from the OLS fit. This will be discussed in more detail when we describe the specific choices made in constructing the Wave 1 emulators for the 7 Galform outputs in the next section. It should be noted that when building an emulator of a complicated function, there are many possible choices to be made (e.g. active variables, regression terms and correlation parameters). Regular consultation with the expert is an important part of this process.

5.1.4 The Wave 1 Emulator

We now describe the construction of the 7 univariate emulators corresponding to the 7 luminosity outputs of interest identified in section 5.1.2. These emulators are used in the first wave of analysis to define the Wave 1 implausibility measures that are required to reduce the input space.

As was discussed in section 5.1.2, the collection of 17 input parameters was split into a group of 8 possibly active parameters and a group of 9 inactive parameters. 993 runs for each of the first 40 sub-volumes were completed from a Latin Hypercube design over the first group of 8 parameters, and these were used to construct the wave 1 emulators. Here the quantity of interest is the mean output over the first 40 subvolumes and writing $f_i^{(j)}(x)$ as the i th output from the j th sub-volume, we define:

$$f_i(x) = \frac{1}{40} \sum_{j=1}^{40} f_i^{(j)}(x). \quad (13)$$

Our approach involves emulating $f_i(x)$ in detail, using only the set of 8 possibly active variables. We then build in the uncertainty due to sampling only 40 sub-volumes from the full 512, and the uncertainty due to the remaining 9 inactive parameters in section 6.1.1. Writing $x_{[B_i]}$ as a vector that spans the space of the 8 possibly active variables, we now assume the following form for the emulator of each $f_i(x_{[B_i]})$ similar to that of equation 12,

$$f_i(x_{[B_i]}) = \sum_j \beta_{ij} g_{ij}(x_{[A_i]}) + u_i(x_{[A_i]}) + w_i(x_{[B_i]}), \quad (14)$$

where the active variables $x_{[A_i]}$ are a subset of the collection of 8 inputs represented by $x_{[B_i]}$. In choosing the active variables the aim is to explain a large amount of the variance of $f_i(x)$ using as few variables as possible. For each of the 7 outputs, the set of 8 inputs was initially reduced by backwards stepwise elimination, starting with a model containing the 8 linear terms in $x_{[B_i]}$. Then individual inputs were discarded in turn based upon the size of their main effect. Before an input would be discarded, a third order polynomial was fitted to see the extent of variance explained with the current set of active variables. It was found that 5 active variables could explain satisfactory amounts of the variance of $f_i(x)$ for each output i (see table 5.1.4), based on the adjusted R^2 of the polynomial fits. In each case, more than 5 variables would yield little extra benefit (compared to the increase in the size of the input space), while less than 5 would lead to substantially worse fits. Once the set of active variables has been determined, the full set of regression terms can be chosen. This was done by forward stepwise selection starting with a model containing the linear terms in the active variables, and adding possible terms from the full 3rd order polynomial in the active variables, using standard stepwise routines in R, based on criterion such as AIC. Note that a reasonable number of runs are required to enable the fitting of a third order polynomial. Here, we are fortunate to have 993 runs of the model which should provide enough degrees of freedom to make over-fitting unlikely, although, even here we need to check our results both against physical intuition and through diagnostic analysis. When the regression terms have

Output	bj 17	bj 21	bj 22.25	K 21	K 22.25	K 24.75	K 25.75
vhotdisk	x	x	x	x	x	x	x
aReheat	x	x	x	x	x	x	x
alphacool		x	x			x	x
vhotburst	x	x	x	x	x	x	x
epsilonStar	x	x		x			
stabledisk			x		x	x	x
alphahot	x			x	x		
yield							
Adj R^2	0.92	0.59	0.70	0.87	0.75	0.72	0.80

Table 2: Wave 1 Active variables and adjusted R^2 for the bj and K luminosity emulator.

been chosen for each output $f_i(x)$, estimates for the $B = \{\beta_{ij}\}$ coefficients can be obtained using Ordinary Least Squares, assuming uncorrelated errors. The results of this procedure, including a list of all model terms and estimates can be found in Appendix A.

As the $u_i(x_{[A_i]})$ represent local deviations from the regression surface we assume that there will be a large correlation between u_i at neighbouring values of the active inputs $x_{[A_i]}$, and specify the following Gaussian covariance structure:

$$\text{Cov}(u_i(x_{[A_i]}), u_i(x'_{[A_i]})) = \sigma_{u_i}^2 \exp(-\|x_{[A_i]} - x'_{[A_i]}\|^2 / \theta_i^2), \quad (15)$$

where $\sigma_{u_i}^2$ is the point variance at any given $x_{[A_i]}$, θ_i is the correlation length parameter that controls the strength of correlation between two separated points in the input space (for points a distance θ apart, the correlation will be exactly $\exp(-1)$), and $\|\cdot\|$ is the Euclidean mean. As the nugget process $w_i(x_{[B_i]})$ represents all the remaining variation in the inactive variables it is often small and we treat it as uncorrelated random noise with $\text{Var}(w_i(x_{[B_i]})) = \sigma_{w_i}^2$. We consider the point variances of these two processes to be proportions of the overall residual variance of the computer model given the emulator trend: σ_i^2 , and write that $\sigma_{u_i}^2 = (1 - w_i)\sigma_i^2$ and $\sigma_{w_i}^2 = w_i\sigma_i^2$ for some small w_i . Various techniques for estimating the correlation length and nugget parameters θ_i and w_i from the data are available (for example variograms, REML); however, these estimation procedures can often be non-robust as the output from a computer model rarely behaves like an actual Gaussian Process. An alternative is to specify the θ_i parameters a priori [5] followed by an approximate assessment of the nugget term w_i , which is the approach we adopt here.

It is possible to provide an approximate interval for the correlation length parameters θ_i , by appealing to the simple heuristic that the regression residuals may be viewed as deriving from a polynomial of order one higher than the fitted polynomial, as they correspond to the first order of terms which are neglected by the regression fit. Here this implies that values of θ_i should be chosen corresponding to the shape of a 4th order polynomial. In such a case, we would not want the correlation length to be greater than the average distance between roots of a 4th order polynomial: approximately 0.25 of the range of the input. Alternatively it can

be argued that there should be positive correlation between outputs at the turning points and the adjacent roots of the polynomial, and that the correlation length must therefore be greater than this distance: approximately 0.125 of the range of the input. While this argument is very rough, it does tend to give more conservative (i.e. smaller) specifications for the correlation length compared to say maximum likelihood or variogram methods. As we have scaled all inputs to the range $[-1, 1]$, this argument suggests that a working estimate of θ_i might lie between 0.25 and 0.5, and we selected the same value for all θ_i of 0.35, based in part on diagnostic checking as we describe in section 5.2.2.

The value of the nugget parameter w_i represents the proportion of residual variance due to the (in this case 3) inactive variables. We obtained a working assessment of w_i by examining the variance explained by the inactive variables for each of the seven outputs, and comparing this to the residual variance from the active variable polynomial fit. These considerations led to a conservative value of 0.2 for w_i acknowledging a reasonable contribution from the inactive variables at each output.

Provided conservative choices are made and are combined with analysis of the emulator diagnostics that we discuss in the next section, such specifications can lead to emulators of sufficient accuracy for the required task, while avoiding the complex and often misleading problem of estimating such parameters from the data alone. It should also be noted that at this stage of the history matching process, we only required a relatively simple emulator in order to make an initial reduction of the input space, while leaving the construction of more detailed emulators to subsequent waves of the analysis.

Emulator construction should be performed in conjunction with physical considerations of the model in question. The emulator should reproduce to a reasonable degree of accuracy, the outputs of the model, and should therefore share the physical features of the model. Careful expert assessment regarding the choice of the active variables and the form of the polynomial fit for each output was made to ensure that the emulators were consistent with insight into the interior workings of the model. For example, consider the polynomial for the first *bj* output given in Appendix A. There are large (negative) contributions from terms involving *vhotdisk* and *alphahot* including a strong interaction between them. Both these parameters are used in the SNe feedback module of the Galform model and increasing either will decrease the luminosity function at the faint end. They are known to interact in the model, and therefore the form of the terms in the polynomial that they feature in makes physical sense.

5.2 Diagnostics

It should be appreciated from the above description that emulator construction is not an exact science. Depending on the shape of the function, building an accurate emulator can be very difficult, especially if a high level of accuracy is required over the whole of the input space. For the case of history matching, where interest lies only in identifying the set of acceptable inputs, such a requirement is not necessary: we only want the emulator to possess a level of accuracy that allows sizable regions of the input space to be discarded with confidence.

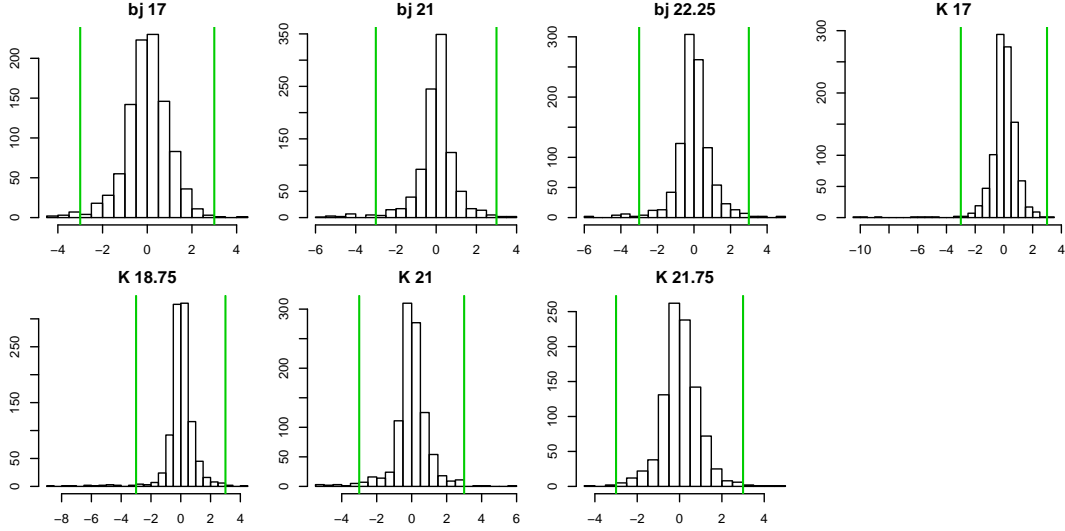


Figure 6: Histogram of the 993 Residuals from the third order polynomial fits for each of the outputs. Vertical lines show the 3 sigma interval.

In order to assess the accuracy of the Wave 1 emulator, diagnostic tests were performed. It should be noted that the nature of the history matching procedure, where we reject input points on an individual point by point basis, implies that we similarly focus on diagnostics on individual points.

5.2.1 Residuals

We first examine the residuals from the polynomial fits described in section 5.1.4 and given in Appendix A. Figure 6 shows histograms of the 993 (standardised) residuals from the polynomial fits for the 7 outputs used in Wave 1. The vertical green lines are at ± 3 sigma, and it can be seen that the majority of points lie within this interval, with an acceptable number of outliers. Figure 7 shows the residual plots. Points where the linear model prediction is closer to the observed data than the actual output are coloured blue. This highlights that although the linear model might be inaccurate for certain inputs, if its predictions are conservative (i.e. closer to the observed data) then we will not accidentally rule out regions of input space that might prove to be acceptable. The red points are the opposite case: here the emulator is giving predictions that are further from the observed data than the actual run output.

As can be seen, several points lie outside of the 3 sigma interval given by the horizontal lines. However, the majority of these points are blue and hence should not cause problems with the history match at this stage (the group of low blue points seen for outputs bj 21, K 17 and K 18.75 are caused by the same group of unphysical runs). There are issues with output 1 for larger values: often the linear model predicts too large a value. However, we checked that there were no runs that gave red points that were outside the 3 sigma level for

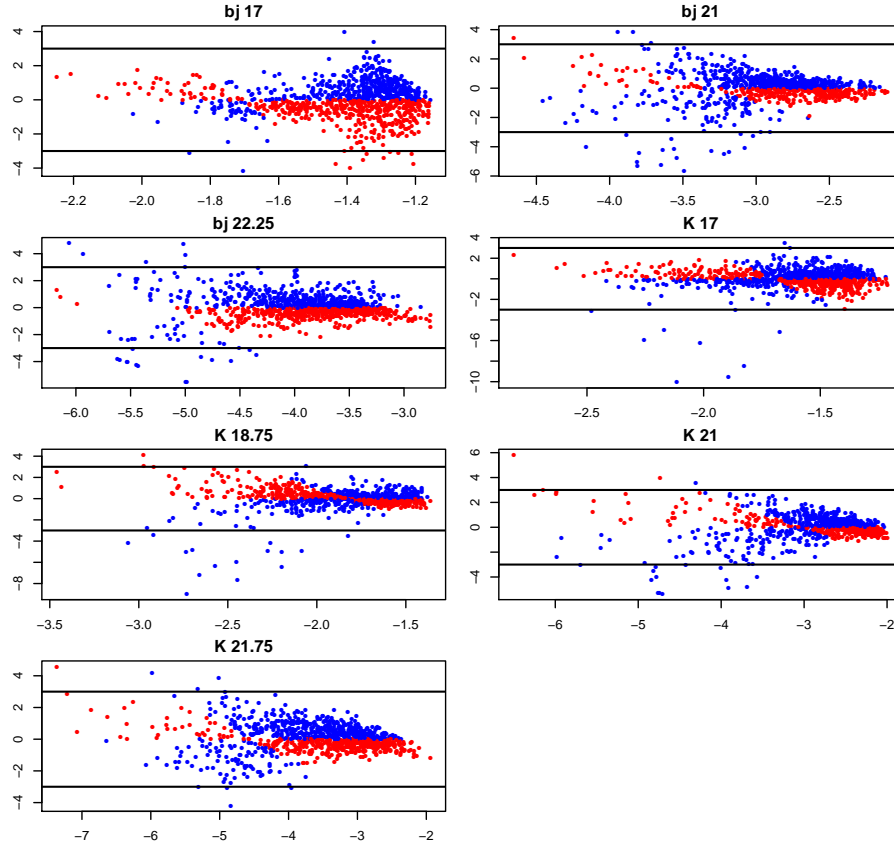


Figure 7: Residuals from the 7 third order polynomial fits, with the x-axis representing the fitted value and the y-axis the standardised residual. Blue points indicate the emulator is predicting the output will be closer to the observed data (a conservative prediction) and red points indicate the opposite. Note that each plot has 993 points.

more than one output. We use implausibility measures that are relatively insensitive to the failings of an individual emulator (see section 7.1), and hence these diagnostics suggest the emulators are sufficiently accurate for our needs. It should be noted that such effects are not unexpected: Galform is a complex function that is not necessarily accurately described by a third order polynomial over all of the input space. Corners of the space where the function decreases rapidly will be missed by the polynomial, but this would only be problematic if this occurs at a part of the input space that is of interest for history matching.

5.2.2 Emulator Prediction Diagnostics

A critical test of an emulator is that of predictive diagnostics. An additional set of 200 diagnostic runs were evaluated over the full input space. The points were chosen by generating several latin hypercube designs, and using the one that had the largest minimum distance to the 993 runs used to construct the emulators. This ensures that the 200 points will provide

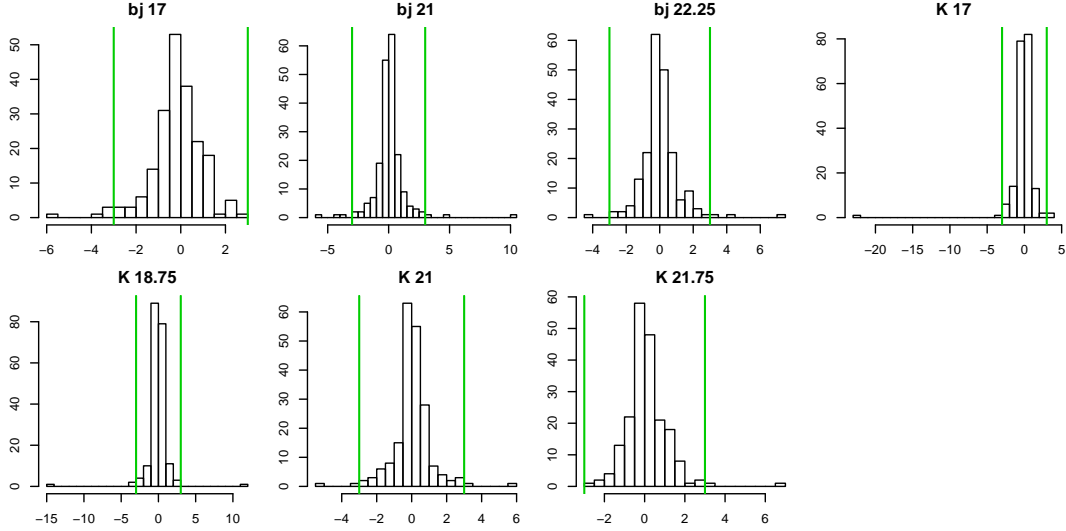


Figure 8: Histograms of Prediction diagnostics for the Wave 1 univariate emulators. Plots show the number of standard deviations the observed output lies from the predicted output: $(f(x) - E(f(x)))/\sqrt{\text{Var}(f(x))}$, with the vertical lines giving 3 sigma intervals.

a satisfactory test of the emulator, as they will not be too close to known model runs.

Here we compare the emulator predictions for the 200 diagnostic runs with the actual outputs, standardised with respect to the emulator variance. Figure 9 shows the histograms of these 200 prediction diagnostics which are calculated using:

$$(f_i(x_{[B_i]}) - E(f_i(x_{[B_i]})))/\sqrt{\text{Var}(f_i(x_{[B_i]}))}, \quad (16)$$

where for each of the 200 runs $x_{[B_i]}$ represents value of the corresponding input, $f_i(x_{[B_i]})$ is the i th output, and $E(f_i(x_{[B_i]}))$ and $\text{Var}(f_i(x_{[B_i]}))$ are the emulator expectation and variance for output i . Figure 8 gives acceptable results: although there are a few outliers, the majority of the 200 diagnostic runs are clearly within the 3 sigma intervals given by the green lines. Figure 8 shows these 200 prediction diagnostics, and, similar to figure 7, the blue points represent conservative emulator predictions, where $E(f_i(x_{[B_i]}))$ lies closer to the observed data than the actual output $f_i(x_{[B_i]})$ itself, and hence for these points will not accidentally rule out input space due to emulator inaccuracy. The red points show the opposite i.e. non-conservative estimates. A similar pattern is seen as in the earlier residual plots: although there are a few places where the emulator does not achieve a high level of accuracy, in most cases it gives conservative predictions and this is acceptable for our purposes. In some places, namely for output 1, the predictions are inaccurate and could lead to the discarding of some parts of the input space incorrectly, but, as was the case for the residuals, none of the diagnostic runs were found to give a non-conservative (red) point larger than 3 sigma in more than one output plot.

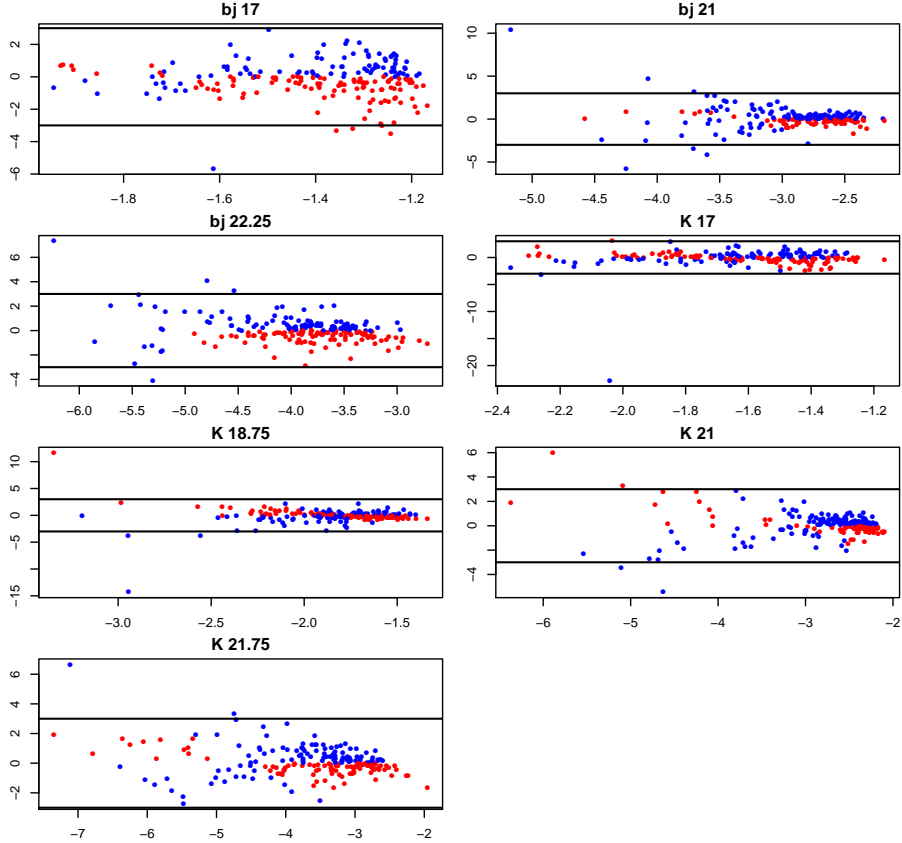


Figure 9: Prediction diagnostics for the Wave 1 univariate emulators. Plots show the number of standard deviations the observed output lies from the predicted output: $(f(x) - E(f(x)))/\sqrt{\text{Var}(f(x))}$, with the blue points indicating where the emulator gives a conservative estimate, and the red points a non-conservative estimate. The horizontal lines show 3 sigma intervals.

6 Quantification of Uncertainty

We now discuss the assessment of all of the remaining uncertainties relevant to linking the Galform Model to the real Universe. These uncertainties can be divided into 2 classes. The first corresponds to the Model Discrepancy ϵ_{md} which describes the possible deficiencies of the model and this has three contributions. These are the uncertainty due to the 9 inactive variables that were not included in the Wave 1 analysis, the effect of the unknown Dark Matter content of the real Universe and finally the structural deficiencies of the Galform model as a whole. The first two contributions can be assessed using additional runs of the model, while the third requires expert assessment.

The second class of uncertainties is that of the observational errors: the luminosity function data has been heavily processed and has several distinct types of error due to uncertainties in the processing, and due to measurement errors in obtaining the original

data.

6.1 Model Discrepancy

We now turn our attention to the link between the model and the system, represented by the model discrepancy. Ideally, Galform would give a perfectly accurate description of all the Galaxies that it attempts to simulate. As is the case with most complex models of physical systems, however, modeling assumptions and approximate solutions to known physical equations imply that Galform’s output will only be an approximation to what would occur in the real Universe. The situation with Galform is more subtle than with many complex models as it does not model specific galaxies that exist within our Universe: instead it simulates around a million galaxies from a ‘possible’ universe that should share statistical properties with our own. The argument is still the same though: these statistical properties will be incorrect due to approximations inherent in the Galform modeling process.

Before meaningful comparisons can be made between the model output and observed data, this discrepancy between the model and the system must be formally addressed. As discussed in section 3.2, we represent the function f (which gives the mean luminosity output over the first 40 sub-volumes) evaluated at the actual system properties x^* as $f^* = f(x^*)$. In order to make meaningful statements about the system, denoted y , in relation to the model, we link the simulator to the system using the *model discrepancy* denoted ϵ_{md} via the equation:

$$y = f^* + \epsilon_{md}, \quad (17)$$

where we consider that ϵ_{md} is uncorrelated with f^* .

The Model Discrepancy term ϵ_{md} links the real system y to the evaluation of the model represented by f^* . The model discrepancy is distinct from other sources of uncertainty in our analysis and we decompose it into three uncorrelated contributions:

$$\epsilon_{md} = \Phi_{IA} + \Phi_{DM} + \Phi_E. \quad (18)$$

where Φ_{IA} represents the discrepancy due to the nine inactive parameters that we were unable to vary in the initial stages of the project, Φ_{DM} is the discrepancy due to the unknown Dark Matter configuration of the real Universe and Φ_E summarises the structural deficiencies of the full Galform model itself. We give details as to the assessment of each of these contributions in the next three sections. It should be noted that quantification of ϵ_{md} is fundamental to our approach as we cannot determine which inputs x are acceptable without such judgements.

6.1.1 Uncertainty Due to Inactive Variables: Φ_{IA}

In section 5.1.4 we introduced the emulator for $f_i(x_{[B_i]})$, which represents the mean output over the first 40 sub-volumes for the function defined over the initial 8 inputs considered, which are spanned by $x_{[B_i]}$. As we were unable to run the Galform model while varying all 17 inputs simultaneously, we could not model the effect of the remaining 9 inactive variables in detail. Therefore, we treat the effect of the 9 variables as initially contributing an extra

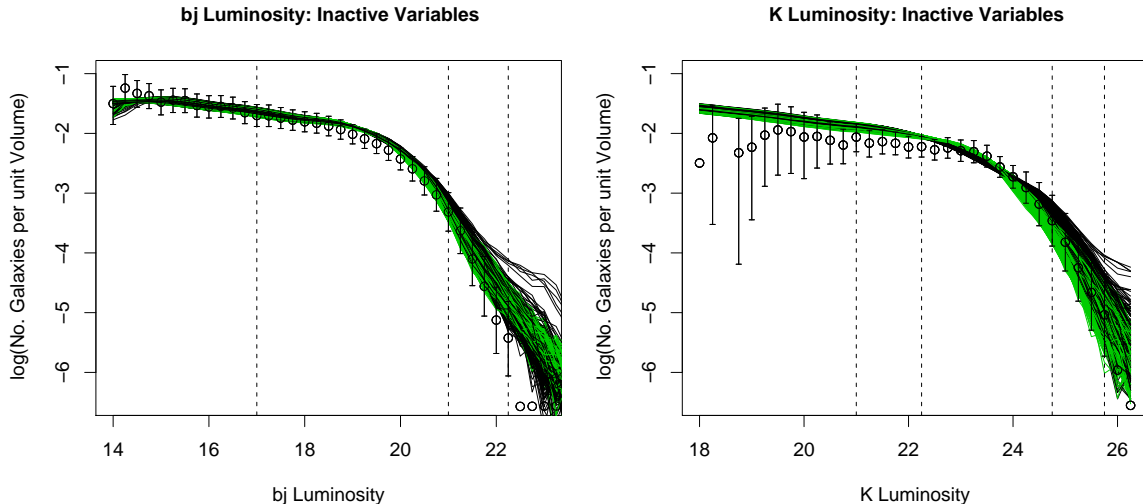


Figure 10: The bj and K luminosity outputs from a sample of 500 runs of the model where only the 9 inactive parameters have been varied (taken from the full 1000 runs shown in figure 2). Green and black lines represent the model output when `tdisk` is off or on respectively, and the observed data and error bars are as in figure 5. It can be seen that varying the inactive parameters causes a small variance in the model output compared to the 8 active parameters (the effects of which are shown in figure 5), and hence justifies the treatment of the 9 inactive parameters as an unstructured model discrepancy term Φ_{IA} . The variance due to the inactive parameters of the bj luminosity function (left panel) can be seen in figure 13 (the light blue line) as a function of luminosity.

term Φ_{IA} to the model discrepancy. (By the time we had reached the fourth wave of our analysis the technical problems that had prevented altering all 17 parameters in the model had been resolved, and we could then model the effects of these extra inputs in more detail.) Note that for the first three waves we are essentially running a reduced model (using only 8 inputs), and therefore must use Φ_{IA} to account for the fact that the Galform model output may not match the observed data due to incorrect settings used for the remaining 9 inputs.

The approximate assessment of Φ_{IA} was performed as follows. We assumed that there would be no overall bias due to the extra 9 inputs and set $E(\Phi_{IA}) = 0$. Recall that these variables have already checked for main effects as discussed in section 5.1.2). Assessing the magnitude of the variance of Φ_{IA} was relatively straightforward as we had performed 1000 runs across the 9 inactive variables (with the original 8 inputs set at their default values) over the first 40 sub-volumes as is described in section 5.1.2. We took the mean of the first 40 sub-volumes for each of these runs, and set the $\text{Var}(\Phi_{IA})$ to be equal to the sample variance of the collection of 1000 means. Note that by making this approximate assessment we are treating as negligible any interactions between the 9 inactive variables and the choice of subvolume, and with the 8 original variables. In figure 10 we show the first 500 out of the set of 1000 runs performed across these 9 input, with the 8 active variables set at the

default value (which corresponds to the cosmologists’ best match: a run which is borderline acceptable according to our matching criteria).

Figure 13 compares the standard deviation of all uncertainties discussed in this section, at every point on the bj luminosity function graph (given in figure 1). The three bj points that were chosen for emulation are given by the black dashed lines. The size of $\sqrt{\text{Var}(\Phi_{IA})}$ for all bj luminosity outputs is shown as the light blue line in figure 13.

Note the similarity between the nugget term denoted $w_i(x_{[B_i]})$ in the Wave 1 emulator of equation (14), which describes the effects of the 3 inactive variables for each output, and the model discrepancy term given by Φ_{IA} . Both are treated as independent of x , have expectation zero and constant variance. Treating these terms in this manner is an initial simplification that makes subsequent calculations far more tractable and allows a straightforward reduction of the input space in the first wave of analysis. In subsequent waves, we model these effects in more detail; for example, in Wave 2 we assume 8 active inputs and hence the $w_i(x_{[B_i]})$ term will be absorbed into the regression and stationary process terms ($u_i(x_{[A_i]})$) of the emulator defined for that wave. This is a typical feature of our approach: the modeling and techniques used are at the minimum level of complexity that will ensure that substantial amounts of input space will be discarded at that wave.

6.1.2 Dark Matter Uncertainty: Φ_{DM}

We now assess the uncertainty due to the difference between the underlying Dark Matter configuration used in each sub-volume run of the Galform model and the unknown Dark Matter configuration that is thought to exist in the real Universe. As is discussed in section 4, the Millenium Simulation provides a detailed description of the evolution of Dark Matter over a large volume, from the beginning of the universe until the current day. This volume is split into 512 sub-volumes each of which can be used by the Galform model. Due to considerations of computational resources, and discussions of principle regarding the magnitude of the variance across individual sub-volumes, it was decided to perform runs using only the first 40 sub-volumes out of the full 512. This choice was also made to facilitate comparison between our study and a previous attempt to find an acceptable match by the cosmologists. While using a larger number of sub-volumes would be more accurate, the extra run time would allow fewer evaluations of points in the input space. As is described in section 5.1.4, we have therefore emulated the mean of the function output over these 40 sub-volumes given by $f_i(x)$. Figure 11 shows the luminosity output from all 40 sub-volumes for two runs of the model (given by the collection of red and blue lines).

The processing of the observational data and associated errors has effectively elevated the data to represent the density of galaxies as measured over a much larger volume of the Universe than is defined by the 512 sub-volumes of the Galform model. We take this volume to be effectively infinite and represent the uncertainty due to analysing the mean of only 40 sub-volumes as the model discrepancy term Φ_{DM} .

We assessed Φ_{DM} by first assuming no overall bias and set $E(\Phi_{DM}) = 0$. We then used the outputs $f_i^{(j)}(x)$ for each of the 40 sub-volumes for the 993 runs performed in Wave 1 to derive an approximate value for the variance of Φ_{DM} as follows. For each of the

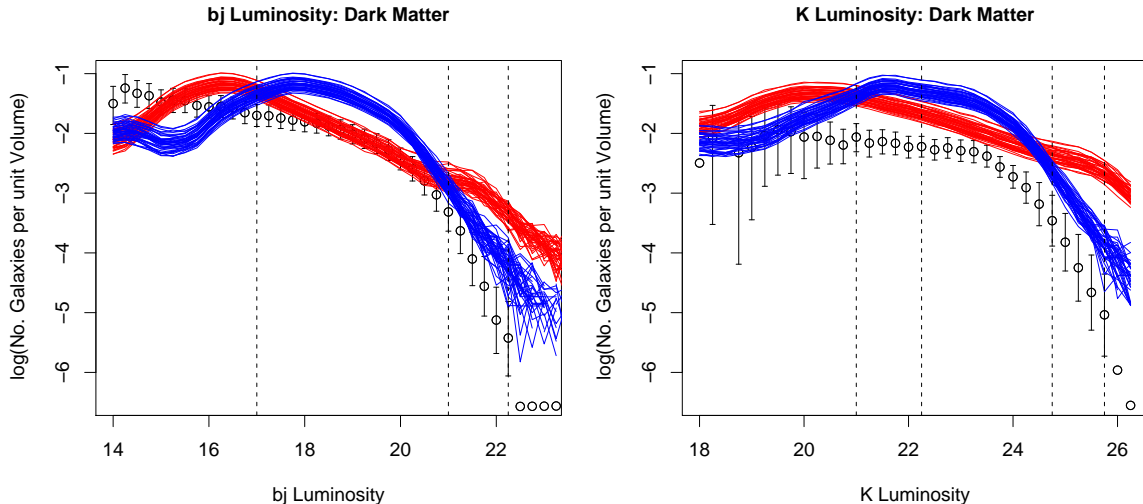


Figure 11: The bj and K luminosity function output of the first 40 sub-volumes of the Dark Matter simulation, for two (blue and red) Wave 1 runs. This source of uncertainty was treated as a model discrepancy term, assumed to have constant variance across all runs.

993 runs we calculated the standard error of the mean output over 40 sub-volumes, and averaged this over all 993 runs. This was done for each of the 7 outputs. While this is a relatively straightforward assessment, given the important simplifying assumption that Φ_{DM} is independent of x , it was felt that this captured the main source of uncertainty without going into detail that would be unwarranted at this stage of the analysis. A more careful treatment would model the outputs of the sub-volumes individually, and we will discuss such approaches using exchangeable computer model techniques in a later section of this work.

In order to check that the first 40 sub-volumes are representative of the full set of 512, we ran a small design of 100 runs at the same x input locations as the first 100 runs of the original Wave 1 design, but now choosing 40 random sub-volumes out of the set of 512 instead of the first 40. We found that the variance across the random 40 sub-volumes was not significantly different from the original 40 and so did not alter the assessment for the $\text{Var}(\Phi_{DM})$ described above.

The size of Φ_{DM} for all bj luminosity outputs (not just the 3 outputs chosen for emulation) is shown as the dark blue line in figure 13. Note that the relative size of Φ_{DM} is small compared to other sources of uncertainty, so that it was considered unnecessary to model its effect in more detail at this stage.

6.1.3 Full Galform Model Discrepancy: Φ_E

As we have identified 7 outputs from the bj and K luminosity functions to be emulated, the model discrepancy ϵ_{md} is a 7 vector, the components of which need to be assessed from expert judgements. In the first wave of our analysis we perform only a univariate analysis of each of

the 7 outputs, hence we required a univariate assessment of each of the components of ϵ_{md} . In wave 2, 3 and 4, multivariate analyses were performed and hence a more detailed multivariate assessment of ϵ_{md} was required. We describe here the full multivariate elicitation.

As we are employing a Bayes Linear analysis, we only require specification of expectations and variances over all quantities of interest. All that is required is a subjective assessment of each value $E(\epsilon_{md})$ and $Var(\epsilon_{md})$, which is still a difficult task.

Expert assessment for beliefs regarding deficiencies of the model was that discrepancy judgements were symmetric in that $E(\epsilon_{md}) = 0$. For the multivariate case, assessment of $Var(\epsilon_{md})$ was required which is now a 7x7 matrix. The structure of this matrix came from Richard's opinion as to the deficiencies of the model as follows.

In the case of Galform there are two major physical defects that can be identified. The first is the possibility that the model has too much (or too little) mass in the simulated universe, possibly due to incorrect choices for the cosmological parameters used in the Millennium simulation (see section 4.3). This would lead to the 7 luminosity outputs all being too high (or too low), and would lead to positive correlation between all outputs in the $Var(\epsilon_{md})$ matrix. The second possible defect is that the model incorrectly calculates the colour of the galaxies, due to inaccurate modeling of stellar populations or dust. This would lead to an apparent increase/decrease in the number of red galaxies and decrease/increase in the number of blue galaxies. This is represented as contributing a smaller negative correlation between the bj and K luminosity outputs. To respect the symmetries of these possible defects, the multivariate Model Discrepancy was parameterised in the following (3+4)x(3+4) block form:

$$Var(\epsilon_{md}) = a^2 \begin{pmatrix} 1 & b & b & c & c & c & c \\ b & 1 & b & c & c & c & c \\ b & b & 1 & c & c & c & c \\ c & c & c & 1 & b & b & b \\ c & c & c & b & 1 & b & b \\ c & c & c & b & b & 1 & b \\ c & c & c & b & b & b & 1 \end{pmatrix} \quad (19)$$

where now a^2 is the univariate variance of the model discrepancy; b is the correlation between outputs of the same luminosity graph (either bj or K luminosity) and c is the cross graph correlation. Making detailed judgements regarding the model discrepancy is still a very difficult task even within a Bayes Linear analysis: while Richard was satisfied with the form of the parameterisation of $Var(\epsilon_{md})$ as given by equation (19), he was understandably cautious about specifying exact quantities for the parameters a , b and c . He was, however, willing to provide the following ranges for a , b and c :

$$3.76 \times 10^{-2} < a < 7.52 \times 10^{-2}, \quad 0.4 < b < 0.8, \quad 0.2 < c < b. \quad (20)$$

This assessment involved examining the difference between Galform and a competing model of similar complexity, consideration of the above possible physical defects to the model, and from his previous years of experience coding and running such galaxy formation models.

After the initial assessment we constructed an elicitation tool in order for Richard to confirm that his specification agreed with his intuition regarding the outputs of the luminosity

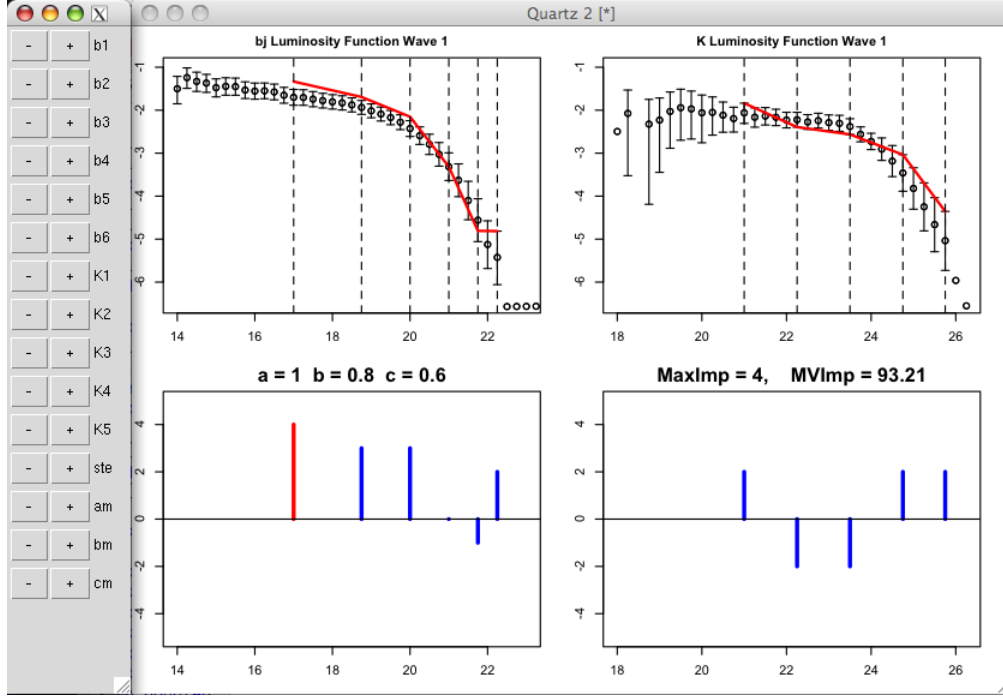


Figure 12: The Elicitation Tool used to confirm the multivariate model discrepancy assessment represented by equations (19) and (20). It allows the expert to construct and adjust fictitious luminosity functions, and to explore the response of the implausibility measures to changes in a , b and c (see section 7).

function. A picture of this elicitation tool is shown in Figure 12, and it possesses the following features. The top two panels of the tool show the bj and K luminosity functions, with observational data points in black, error bars representing all uncertainties, dotted lines giving the 11 outputs of interest (additional outputs were used in later waves), and constructed (or fictitious) luminosity model output given by the red lines. The constructed model output lines can be controlled by the user with the first 11 controls on the left (grey) panel labelled b1-b6 and K1-K5. These controls allow independent adjustment of each of the 11 outputs (by varying increments controlled by the 'ste' button) in order to represent any possible luminosity function output. The bottom two panels show the number of standard deviations that each output is from the observed data, with the furthest away in red. Above the bottom right panel the values of the two implausibility measures 'MaxImp' ($I_M(x)$) and 'MVImp' ($I(x)$) are given, calculated using the current constructed luminosity output (see section 7.1 for definitions of these measures). The user can specify when starting the tool, which uncertainties they want to be considered in the implausibility calculation (e.g. use all observational and model discrepancy uncertainties, or purely the Φ_E component).

This elicitation tool allows the user to experiment with various possible luminosity functions and see the corresponding values for the two implausibility functions $I_M(x)$ and $I(x)$.

Most importantly, the values of the multivariate model discrepancy parameters a , b and c can be controlled by the ‘am’, ‘bm’ and ‘cm’ buttons, with current values shown above the bottom left panel (a is given in terms of multiples of Richard’s original assessment). This allowed Richard to experiment with different specifications of a , b and c and to see the response of the implausibility measures. This is useful for the expert to get a feel for the behaviour of a multivariate implausibility measure, understand the ramifications of the assumed structure of $\text{Var}(\epsilon_{md})$ and also to check that intuitively acceptable runs would not be ruled out by the current specification.

Obviously it is possible to build in far more structure into $\text{Var}(\epsilon_{md})$ if required. The aim here was to account for the main sources of model discrepancy, while maintaining a relatively simple structure of the $\text{Var}(\epsilon_{md})$, as the more detailed the structure, the more difficult eliciting expert information becomes. However, note the relative ease of specifying useful high-level statements using expectation as primitive, as compared to the corresponding effort for a fully probabilistic analysis.

As we have ranges for the parameters a , b and c we will incorporate this into our analysis when we reduce the input space using various implausibility measures. Effectively we perform a sensitivity analysis, and rule out parts of the input space only if they fail certain implausibility cutoffs for all values of a , b and c within the above ranges. This will be discussed further in later sections.

6.2 Observational Errors

The final set of uncertainties that need to be considered before we can proceed with the first wave of History Matching are those relating to the Observational Errors. This is a complex topic as the generation of the b_j and K luminosity function data, shown in figure 1, is an extremely intricate task. Sky Surveys are performed where telescopes sweep across sections of the sky, measuring among other things the brightness and colour of tens of thousands of galaxies. The maximum distance at which galaxies can be accurately measured is a function of the brightness of the galaxy in question: bright galaxies can be seen at large distances, while dim ones can only be detected if they are relatively close to our own Milky Way. This data is then processed using information from deeper sky surveys, large scale simulations of the Dark Matter content of the Universe, and combined with other experimental and theoretical knowledge related to the evolution of the Universe and the galaxies within it. Such processing leads to the data shown in figure 1, which represents the best measurement of the b_j and K luminosity functions of the Universe itself, namely the average number of galaxies per unit volume, per luminosity for the whole observable Universe. Such measurements and subsequent transformation of the raw data generate interesting types of error which we describe below.

First we need formally to link the real system (the actual luminosity function of the Universe) to the measurements described above. As described in section 3.2 we are not able to measure the real system y directly, we link the system y with the observational

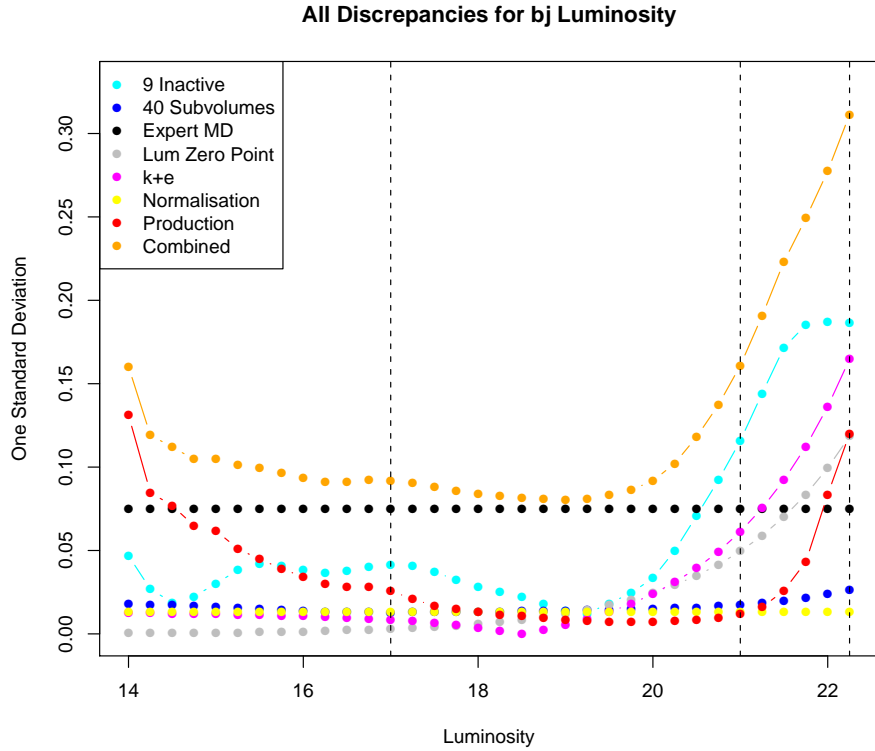


Figure 13: Plot showing all the contributions from the various sources of uncertainty for the full range of the bj luminosity function (the x-axis is the same as figure 1). The vertical lines show the location of the three bj outputs chosen for emulation in Wave 1, and the coloured lines give one standard deviation for each of the contributions. The orange line represents one standard deviation of the total uncertainty due to all contributions, and it is this value that is used in all bj luminosity plots such as figure 1.

measurements z via the equation:

$$z = y + \epsilon_{obs}, \quad (21)$$

where ϵ_{obs} represent the observational errors and has $E(\epsilon_{obs}) = 0$. The $\text{Var}(\epsilon_{obs})$ has four contributions related to the measurements and to the processing of the data described above. These are:

The Luminosity Zero Point Error - this is derived from the difficulty of defining the Luminosity Zero Point: that is the point on the x-axis of the luminosity graph (see figure 1) corresponding to a galaxy of ‘zero’ brightness. This results in a correlated error on every output point, the relative magnitude of which can be seen in figure 13 as the grey line.

The k+e error - a perfectly correlated error on all output points due to a correction made to compensate for two important effects. (i) Several of the galaxies measured are so

far away that the light we see from them was emitted billions of years ago. The luminosity function is defined in terms of the luminosities of the galaxies at the present day and so corrections are made (using models of galaxy evolution) for viewing these distant galaxies in the past. (ii) Galaxies at large distances are moving away from Earth at high speeds due to the expansion of the Universe and their light emitted is therefore redshifted. Galaxies might appear to have, for example, high K-band (red) luminosity when actually this light comes from a shorter (bluer) part of their emission spectrum. Corrections are made for this using models of the complete emission spectra of various galaxy types. The $k + e$ error is shown as a purple line in figure 13.

The Normalisation Error - The data on galaxies comes from measurements made in our local vicinity and it is possible that we live in a relatively under/over populated part of the Universe. Assessments made of the luminosity function would therefore be affected and the Normalisation error is calculated to account for this uncertainty. This involves using large simulations of the Dark matter content of a possible universe combined with theoretical understanding of the variation in mass density in the Universe on large scales. Shown as the yellow line in figure 13.

Galaxy Production Error - Bright galaxies can be measured up to reasonably large distances from our Milky Way, whether faint galaxies can only be measured at relatively close distances as they are fainter and therefore harder to detect. This error represents the uncertainty due to sampling galaxies over volumes that depend on their brightness, and is constructed using theoretical considerations which include assumptions regarding the shape of the mean luminosity function over the whole Universe. Shown as the red line in figure 13.

It is clear that significant contributions to the observational errors come from uncertainties related to the processing of the data (i.e. the $k + e$, Normalisation and Production Errors). These are distinct from measurement errors and are derived from complex theoretical and modeling uncertainties, and hence could be referred to as model discrepancy terms as opposed to observational errors. However, the calculations involved in determining these errors are intricate and rely upon specialist knowledge of Astronomy, and although it would be desirable to disentangle some of these errors, due to time constraints it was felt that this was unnecessary at the current stage of the collaboration.

7 First Wave History Match

In this section we describe the procedure used to perform the first wave of History Matching.

7.1 Implausibility Measures

We want to learn about the values of x that will give rise to acceptable matches between model output and observed data, and hence identify the set of all possible x values \mathcal{X}^* . The nature of the set \mathcal{X}^* will depend upon the structure and magnitude of all relevant uncertainties, as well as the behaviour of the function $f(x)$ which we do not know exactly. Following the discussion in section 3.5, we define simple Implausibility Measures over the

input space. The principle behind such measures is that we would expect values of x for which the difference between the expectation of the function $f(x)$ and the observed data z is large, to be highly unlikely to satisfy our matching criteria. Here, large means with respect to all the relevant uncertainties due to Emulator Variance, Model Discrepancy and Observational Errors. Therefore we define the following Univariate Implausibility Measure:

$$I_{(i)}^2(x) = |E(f_i(x)) - z_i|^2 / \text{Var}(E(f_i(x)) - z_i), \quad (22)$$

which now becomes (using equations (1) and (2)):

$$I_{(i)}^2(x) = |E(f_i(x)) - z_i|^2 / (\text{Var}(f_i(x)) + \text{Var}(\epsilon_{md:i}) + \text{Var}(\epsilon_{obs:i})), \quad (23)$$

where $E(f_i(x))$ and $\text{Var}(f_i(x))$ are the emulator expectation and variance for component i of f , z_i are the corresponding observed data and $\epsilon_{md:i}$ and $\epsilon_{obs:i}$ are the (univariate) Model Discrepancy and Observational Error for component i . Such simple, quick to evaluate measures are of great use in computer model analysis as high values of $I_i(x)$ imply that evaluating the Galform function using inputs x is unlikely to yield an acceptable match between the model output and the observational data, and suggest that these values should perhaps be discarded from consideration. Note that $I_i(x)$ can give a low value for two possible reasons: either we expect that evaluating the function $f(x)$ at x will produce an output that is close to the observations (if $\text{Var}(f(x))$ is low), or because we are uncertain about the output of $f(x)$ at this point (due to $\text{Var}(f(x))$ being high). Therefore low values of the Implausibility Measure automatically suggest values of x that it would be desirable to use for future runs of the Galform model, as at these locations we are either likely to obtain good matches to the outputs, or we will learn about the function $f(x)$ in regions where previously its behaviour was sufficiently uncertain that we were unsure whether an acceptable match could be produced. In this way, the Implausibility Measure can be seen as a simple tool to generate a second stage design, a strategy that will be discussed in section 7.3.

Various Implausibility Measures can be defined, either from combinations of the 7 univariate measures, or by constructing multivariate versions using subsets or all of the 7 outputs. We have used four main types of measure in this case study (two of which are used in later waves). The simplest of these is obtained by maximizing over the 7 outputs and we define the Maximum Implausibility Measure $I_M(x)$ as:

$$I_M(x) = \max_i I_{(i)}(x). \quad (24)$$

This measure is used in later waves of our analysis and it represents a major part of the definition of an acceptable match. It is, however, sensitive to problems concerning the inaccuracies of individual emulators, and so we define the Second and Third Maximum Implausibility Measures $I_{2M}(x)$ and $I_{3M}(x)$ as:

$$I_{2M}(x) = \max_i (\{I_{(i)}(x)\} \setminus I_M(x)), \quad (25)$$

$$I_{3M}(x) = \max_i (\{I_{(i)}(x)\} \setminus \{I_M(x), I_{2M}(x)\}), \quad (26)$$

that is defining $I_{2M}(x)$ and $I_{3M}(x)$ to be the second and third highest value out of the set of univariate measures $I_i(x)$ respectively. These were used in wave 1 one as they were thought to be relatively safe measures in that they were less sensitive to the possibility that one of the emulators was inaccurate. Note that in equations (24), (25) and (26) the index i runs over all outputs that are being considered at that Wave: for example, in later Waves, 11 outputs are used in the analysis.

In later waves we will use a Multivariate Implausibility measure defined as:

$$I^2(x) = (\mathbf{E}(f(x)) - z)^T \text{Var}(\mathbf{E}(f(x)) - z)^{-1} (\mathbf{E}(f(x)) - z), \quad (27)$$

which becomes (using equations (1) and (2)):

$$I^2(x) = (\mathbf{E}(f(x)) - z)^T (\text{Var}(f(x)) + \text{Var}(\epsilon_{md}) + \text{Var}(\epsilon_{obs}))^{-1} (\mathbf{E}(f(x)) - z). \quad (28)$$

Again, large values of $I(x)$ imply that we would be unlikely to obtain a good match between model output and observed data were we to run the model at input x . $I(x)$ is a useful measure to consider as it captures the shape of the luminosity function output. It will allow the discarding of inputs corresponding to runs that satisfy the univariate matching criteria and hence are close to the data points, but that have an unphysical shape in either bj or K luminosity function.

7.2 History Matching via Implausibility

History Matching is the process of Identifying the set \mathcal{X}^* , that is the set of points that would give acceptable matches between model output and observational data. Identifying \mathcal{X}^* is a difficult task as often it represents a complicated object in a high dimensional space. \mathcal{X}^* could also be comprised of disconnected volumes, which could even possess non-trivial topology. In many applications \mathcal{X}^* occupies an extremely small fraction of the original input space, with large volumes of input space leading to very poor matches to the observed data.

We employ a relatively straightforward iterative technique where the Implausibility Measures are used to perform the History Matching process. The basic strategy is based around discarding values of x that are highly unlikely to yield acceptable matches between model output and observational data. This is done by applying a cutoff on the Implausibility Measures defined in section 7.1. As the Implausibility Measures are constructed using the emulator, they are fast to evaluate and therefore we can efficiently identify values of x that will be discarded. For example, in Wave 1 we use both the second and third maximum Implausibility Measures $I_{2M}(x)$ and $I_{3M}(x)$ defined in section 7.1 to discard values of x that do not satisfy both:

$$I_{2M}(x) < I_{cut2} \quad \text{and} \quad I_{3M}(x) < I_{cut3}, \quad (29)$$

where I_{cut2} and I_{cut3} are the corresponding implausibility cutoffs.

The choices made for the individual cutoffs come from a combination of examination of diagnostics (such as shown in figure 14), consideration of the amount of space cut out, and unimodality arguments which are employed as follows. Regarding the size of the individual

univariate Implausibility Measures $I_{(i)}(x)$, if we consider that for fixed x the appropriate distribution of $(E(f_i(x^*)) - z)$ is unimodal, then we can use the 3σ rule [15] which implies that if $x = x^*$, then $I_{(i)}(x) < 3$ with a probability of greater than 0.95. Values higher than 3 would suggest that the point x could be discarded. This is a general result, and applies even if the appropriate distribution is asymmetric. We need to specify values for I_{cut2} and I_{cut3} , and while the unimodal argument suggests using cutoffs of 3 or higher (depending on the correlation between outputs), consideration of figure 14 shows that this might be unnecessarily conservative. In response to this we choose cutoffs of $I_{cut2} = 2.7$ and $I_{cut3} = 2.3$ (shown as vertical lines in figure 14), recognising the fact that we want to balance a conservative cutoff with the amount of space that can be removed at Wave 1. These cutoffs resulted in approximately 85.1 percent of the input space being ruled out due to the Wave 1 analysis.

Figure 14 shows diagnostic plots regarding the choice of cutoffs I_{cut2} and I_{cut3} . It shows the maximum data implausibility $I_M^{data}(x)$ (that is the implausibility evaluated at a known run, given by equation (9)) across the 7 outputs for the 200 diagnostic runs (y-axis), against $I_{2M}(x)$ (left panel) and $I_{3M}(x)$ (right panel) that are used to reduce the input space. The vertical lines are the cutoffs that will be imposed, and it can be seen that no points to the right of these lines (in red) actually yielded an acceptable match to the observed data (a possible criterion of an acceptable fit is given by the horizontal line representing a 2 sigma boundary). These diagnostics are therefore consistent with the statement that the space cut out in Wave 1 does not contain any inputs of interest.

In figure 15 we show various 2-dimensional projections (top 3 panels) of values of the Implausibility Measures, with red areas representing high implausibility and green areas low, which were constructed as follows. For each plot we evaluated the emulator at a set of inputs specifically designed to produce a 2-dimensional projection in the appropriate input plane. For example, in the top left panel the projection is in the vhotdisk - alphareheat plane, and the emulator was evaluated on a (2d grid)x(5d latin hypercube) design, where the 2d grid was over the vhotdisk - alphareheat plane (and of size 15^2) while the latin hypercube was defined over the remaining 5 active inputs at Wave 1 (and was of size 1500). For each point on the grid, we then minimised the implausibility over the corresponding 1500 points at that grid location, the results of which provide the plots shown. This allows the following interpretation: a red area in one of these implausibility projection plots implies that even given all relevant uncertainties, and all possible choices for the other input parameters, it is highly unlikely that an acceptable match will be found at this point in the vhotdisk - alphareheat plane (for example).

The bottom 3 panels of figure 15 show depth projection plots: these are constructed by calculating at each grid point, the fraction of the corresponding 1500 points of the latin hypercube that survive the implausibility cutoffs, given by equation 29. This gives information as to the ‘optical depth’ of the the 7 dimensional non-implausible volume when observed in a direction perpendicular to the vhotdisk - alphareheat plane (for example). They provide complimentary information to the implausibility projections. Consider the middle top and bottom panels of figure 15, where the implausibility projection (top panel) shows that

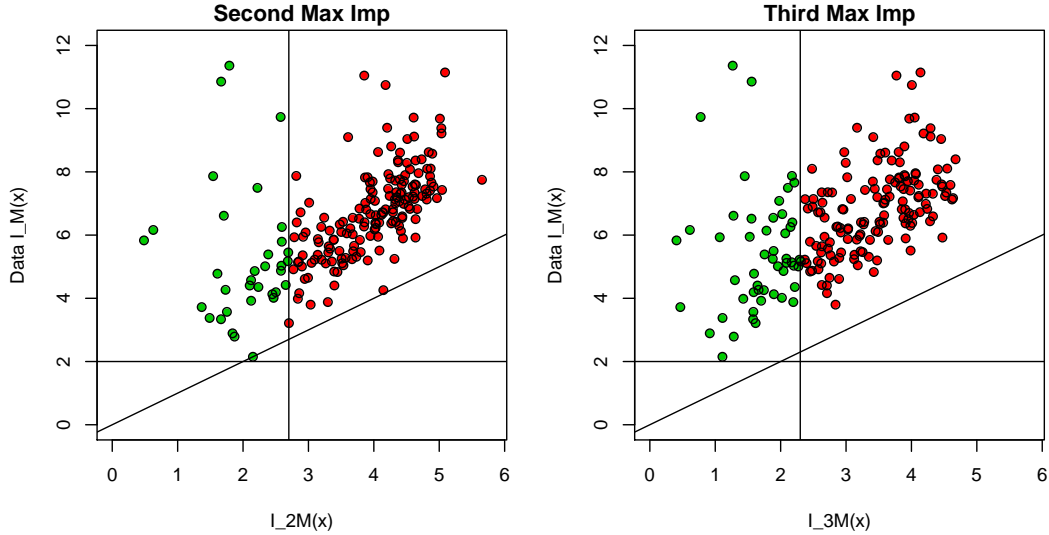


Figure 14: Implausibility diagnostics for the Wave 1 univariate emulators. Plots show ‘maximum data implausibility’ which is defined to be $I_M(x)$ evaluated using known runs and hence involves no emulator uncertainty, against the implausibility measures $I_{2M}(x)$ (left panel) and $I_{3M}(x)$ (right panel) which are calculated using the emulator. The vertical lines show the cutoffs imposed at this Wave, with the red points belonging to parts of the input space deemed implausible. It can be seen that the cutoffs chosen seem relatively safe in that no red (discarded) points were found to actually be acceptable runs (in which case they would be close to the horizontal line).

non-implausible choices of alphareheat and alphacool exist over much of the alphareheat-alphacool plane. The depth plot demonstrates that the majority of the non-implausible volume is found at low values of alphareheat.

These images give physical insights into the nature of the Galform model: in the top right panel of figure 15 we see that simultaneously low values of both vhotdisk and alphahot are ruled out, and that high values of both these parameters are possibly preferred. These parameters are involved in the same Galform module: that of Feedback from Supernovae (see equation (11) and section 4.3), and increasing their size should increase the amount of material expelled from certain galaxies as opposed to being used to form stars. This will reduce the luminosity function at the faint end, and, as most of the Wave 1 runs are higher than the observed data, it makes physical sense that parameter choices that lower the luminosity function will be preferred. These physical features can be seen in the polynomial terms for the outputs bj 17 and K 21 (which are at the faint end of the luminosity function), specifically the large and negative coefficients for the vhotdisk, alphahot and their interaction terms. The Wave 1 emulators are quite approximate, so there is a limit as to the physical insight they, and the corresponding implausibility measures, can provide. We will discuss the physical results of the full analysis in more detail in later sections.

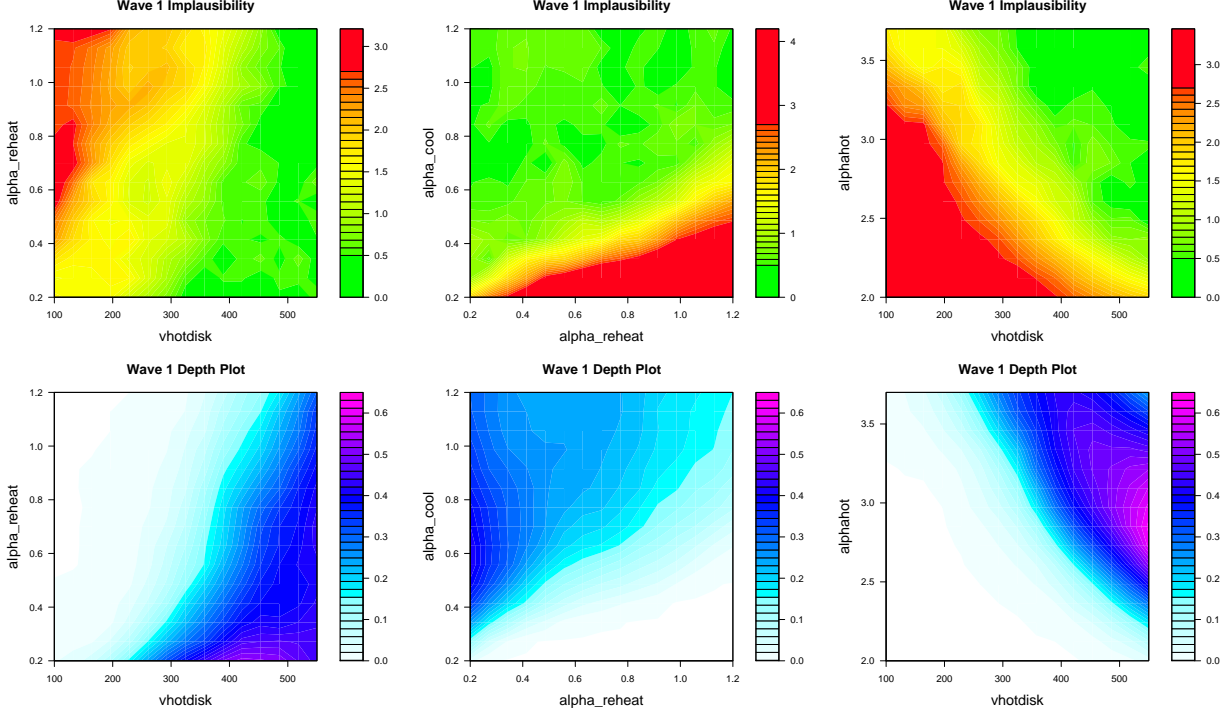


Figure 15: The top three panels give Wave 1 implausibility projection plots: the red region indicates high implausibility where input points will be discarded. Note that the yellow and green regions occupy only 15% of the input space (the non-implausible region), even though they take up much larger areas of the 2-dimensional projection. The bottom three panels give the depth plots, showing the areas where there are lots of choices of the other 5 active inputs that give low implausibility.

Equation (29) defines a volume of input space that we refer to as non-implausible after Wave 1 and denote \mathcal{X}_1 . This non-implausible volume should, hopefully, contain the set \mathcal{X}^* . In the first wave of the analysis which we are describing here, \mathcal{X}_1 will be substantially larger than \mathcal{X}^* . This is because it will contain many values of x that only satisfy the implausibility cutoff given by equation (29) because of a substantial emulator variance $\text{Var}(f(x))$ (see equation (23)). If the emulator was sufficiently accurate over the whole of the input space that $\text{Var}(f(x))$ was small compared to the Model Discrepancy and the Observational Error variances, then the non-implausible volume defined by \mathcal{X}_1 would be comparable to \mathcal{X}^* and the History Match would be complete. However, to construct such an accurate emulator would require an infeasible number of runs of the model. Even if such a large number of runs were possible, it would be an extremely inefficient method: we do not need the emulator to be highly accurate in regions of the input space where the outputs of the model are clearly very different from the observed data.

This is the main motivation for our iterative approach. In each wave we design a set of runs over the current non-implausible volume, emulate using these runs, calculate the

implausibility measure and impose a cutoff to define a new (smaller) non-implausible volume. This is referred to as refocusing and is discussed in the next section.

7.3 Refocusing

The Wave 1 History matching process defines a non-implausible region of input parameter space \mathcal{X}_1 . At this stage we cannot reduce space further due to the various uncertainties involved, one of which is the Wave 1 emulator uncertainty. The other uncertainties that contribute to \mathcal{X}_1 either cannot be reduced (e.g the observational errors, pure expert assessed model discrepancy), or are hard to reduce further (the uncertainty due to the 9 inactive parameters). On the other hand, the emulator uncertainty can be partially resolved by performing more runs over the non-implausible volume \mathcal{X}_1 . Using these runs we can construct a new, more accurate emulator and use this to define a smaller non-implausible volume \mathcal{X}_2 such that $\mathcal{X}^* \subset \mathcal{X}_2 \subset \mathcal{X}_1$. This process is referred to as refocusing and is a major part of our analysis of Galform.

We employ the refocusing technique iteratively as follows. At each iteration or Wave:

1. A design for a set of runs over the current non-implausible volume \mathcal{X}_i is created, using a latin hypercube design with a rejection strategy based on the current implausibility measures.
2. These runs are used to construct a more accurate emulator defined only over the current non-implausible volume \mathcal{X}_i .
3. The implausibility measures are then recalculated over \mathcal{X}_i , using the new emulator.
4. Cutoffs are imposed on the Implausibility measures and this defines a new, smaller non-implausible volume \mathcal{X}_{i+1} such that $\mathcal{X}^* \subset \mathcal{X}_{i+1} \subset \mathcal{X}_i$.
5. Unless the emulator variance is now small in comparison to the other sources of uncertainty, return to step 1.

As we progress through each iteration the emulator at each wave will become more and more accurate, but will only be defined over the previous non-implausible volume given in the previous wave. We expect this improvement in the accuracy of the emulator for several reasons. As we have reduced the size of the input space and have effectively zoomed in on a smaller part of the function, we expect the function to be smoother and to be more easily approximated by a third order polynomial. Hence the regression terms in the emulator equation (14) should produce a better fit, with reduced residual variances for all outputs. Due to the increased density of runs over \mathcal{X}_i compared to previous waves, the stationary process term $u_i(x_{[A_i]})$ (updated by the new runs) will be more accurate and have lower variance as the point $x_{[A_i]}$ will be in general closer to known evaluation outputs.

Another major improvement in the emulators comes from identifying a larger set of active variables. Cutting down the input space also means that the ranges of the function outputs are reduced. Dominant inputs that previously had large effects on the outputs have likewise

been constrained, and their effects lessened. This results in it being easier to identify more active variables that were previously masked by a small number of dominant variables. This is especially important in Wave 4 as it was at this point that we were able to perform function evaluations across all 17 inputs simultaneously. Increasing the number of active variables allows more of the function's structure to be modeled by the third order polynomials, and has the effect of reducing the nugget term $w_i(x_{[B_i]})$ (and in Wave 4, the $\text{Var}(\Phi_{IA})$ term).

As the input space is reduced, it not only becomes easier to accurately emulate existing outputs but also to emulate outputs that were not considered in previous waves. Outputs may not have been considered previously because they were either difficult to emulate, or because they were not informative and hence could not be used to remove a significant part of the input space (compared to other outputs). For example, in Wave 1 only seven outputs were considered, as shown in figure 5, while in later waves a total of 11 outputs were used. We now go on to discuss the analysis and results of Waves 2 to 4.

Here we give the provisional plan of the last 3 sections of the Case Study.

8 Analysis of Waves 2 - 4

We will discuss the analysis used in Waves 2, 3 and 4. This will include briefly describing the design of runs over the non-implausible regions, the choosing of larger sets of active variables and the subsequent construction of more accurate emulators (along with suitable emulator diagnostics).

The use of the multivariate implausibility measure to reduce space further will be discussed, along with imprecise considerations regarding the parameters a , b , and c of the model discrepancy matrix $\text{Var}(\Phi_E)$ and their relation to the decisions made when discarding inputs. Various tables and plots will be presented showing the increase in emulator accuracy over each wave, the space cutout at each wave, the progression of implausibility plots from Waves 1-4, and the physical insight that this generates.

9 Results of Wave 4 and 5 - Visualisation

The Wave 4 emulator gives a detailed form for the non-implausible region of input parameter space. Visualising such a region is a difficult task as it can possess a complicated shape in a high dimensional space. In this section we will discuss various visualisation strategies including efficient emulator design for the purposes of creating suitable 2d and 3d projection plots, relevant calculational simplifications, and the use of the emulator polynomials to perform a fast screening of the input space. We will present a number of 2d and 3d plots in order to understand the features of the non-implausible region, and go on to discuss in detail the connection to the Galform model and the subsequent physical insight that has been gained. We will also discuss interactive tools for viewing high dimensional objects.

Using the Wave 4 emulator we performed a final set of runs to obtain a large number of acceptable runs for use by the cosmologists. We will present these, showing that the non-implausible volume is now indeed populated by high numbers of acceptable runs, and discuss stopping criteria.

10 Conclusion

Here we will present our conclusions, focussing upon the physical understanding that such an analysis provides. We will also summarise the problems faced with such a project, and describe possible improvements to the analysis that we have not incorporated. As this collaboration is an ongoing process, we will discuss future directions of research.

References

- [1] Carlton M. Baugh. A primer on hierarchical galaxy formation: the semi-analytical approach. *Rept. Prog. Phys.*, 69:3101–3156, 2006.
- [2] R G Bower and A J Benson et.al. The broken hierarchy of galaxy formation. *Mon.Not.Roy.Astron.Soc.*, 370:645–655, 2006.
- [3] Shaun Cole et al. The 2dF Galaxy Redshift Survey: Near Infrared Galaxy Luminosity Functions. *Mon. Not. Roy. Astron. Soc.*, 326:255, 2001.
- [4] S Conti, J P Gosling, J E Oakley, and A O’Hagan. Gaussian process emulation of dynamic computer codes. To be published, 2009.
- [5] P S Craig, M Goldstein, A H Seheult, and J A Smith. Bayes linear strategies for history matching of hydrocarbon reservoirs. In J M Bernardo, J O Berger, A P Dawid, and A F M Smith, editors, *Bayesian Statistics 5*, pages 69–95. Clarendon Press, Oxford, UK, 1996.
- [6] P S Craig, M Goldstein, A H Seheult, and J A Smith. Pressure matching for hydrocarbon reservoirs: a case study in the use of bayes linear strategies for large computer experiments. In C Gatsonis, J S Hodges, R E Kass, R McCulloch, P Rossi, and N D Singpurwalla, editors, *Case Studies in Bayesian Statistics*, volume 3, pages 36–93. Springer-Verlag, New York, 1997.
- [7] C Currin, T Mitchell, M Morris, and D Ylvisaker. Bayesian prediction of deterministic functions with applications to the design and analysis of computer experiments. *Journal of the American Statistical Association*, 86(416):953–963, 1991.
- [8] B De Finetti. *Theory of Probability*, volume 1. Wiley, London, 1974.
- [9] B De Finetti. *Theory of Probability*, volume 2. Wiley, London, 1975.
- [10] M Goldstein and J C Rougier. Reified bayesian modelling and inference for physical systems (with discussion). *Journal of Statistical Planning and Inference*, 139(3):1221–1239, 2009.
- [11] M Goldstein and D A Wooff. *Bayes Linear Statistics: Theory and Methods*. Wiley, Chichester, 2007.
- [12] D Higdon, M Kennedy, J C Cavendish, J A Cafeo, and R D Ryne. Combining field data and computer simulations for calibration and prediction. *SIAM Journal on Scientific Computing*, 26(2):448–466, 2004.
- [13] J Oakley and A O’Hagan. Bayesian inference for the uncertainty distribution of computer model outputs. *Biometrika*, 89(4):769–784, 2002.

- [14] A O'Hagan. Bayesian analysis of computer code outputs: A tutorial. *Reliability Engineering and System Safety*, 91:1290–1300, 2006.
- [15] F Pukelsheim. The three sigma rule. *The American Statistician*, 48:88–91, 1994.
- [16] J Sacks, W J Welch, T J Mitchell, and H P Wynn. Design and analysis of computer experiments. *Statistical Science*, 4(4):409–435, 1989.
- [17] T J Santner, B J Williams, and W I Notz. *The Design and Analysis of Computer Experiments*. Springer-Verlag, New York, 2003.
- [18] D. N. Spergel et al. First Year Wilkinson Microwave Anisotropy Probe (WMAP) Observations: Determination of Cosmological Parameters. *Astrophys. J. Suppl.*, 148:175–194, 2003.
- [19] Volker Springel et al. Simulating the joint evolution of quasars, galaxies and their large-scale distribution. *Nature*, 435:629–636, 2005.

A Wave 1 Polynomials

	Estimate	Std. Error
(Intercept)	-1.57	0.01
vhotdisk	-0.28	0.01
alphahot	-0.20	0.01
alpha_reheat	0.08	0.01
vhotburst	-0.08	0.01
epsilon_Star	-0.00	0.01
I(alpha_reheat^2)	-0.07	0.01
I(vhotdisk^2)	-0.06	0.01
I(epsilon_Star^3)	0.09	0.02
I(epsilon_Star^2)	-0.04	0.01
I(alpha_reheat^3)	0.06	0.02
I(alphahot^2)	-0.03	0.01
I(vhotburst^2)	0.02	0.01
vhotdisk:alphahot	-0.21	0.01
vhotdisk:alpha_reheat	0.11	0.01
vhotburst:epsilon_Star	-0.08	0.01
vhotdisk:epsilon_Star	0.08	0.01
alphahot:alpha_reheat	0.07	0.01
vhotdisk:vhotburst	0.03	0.01
alphahot:epsilon_Star	0.03	0.01
vhotdisk:I(alpha_reheat^2)	-0.06	0.02
alphahot:I(vhotdisk^2)	-0.05	0.02
alphahot:I(alpha_reheat^2)	-0.04	0.02
alphahot:vhotburst	-0.02	0.01
epsilon_Star:I(vhotdisk^2)	-0.03	0.02
vhotburst:I(epsilon_Star^2)	0.03	0.02
vhotburst:I(alpha_reheat^2)	-0.02	0.02
alpha_reheat:vhotburst	0.02	0.01
vhotdisk:alphahot:alpha_reheat	0.04	0.02
vhotdisk:alphahot:epsilon_Star	0.04	0.02
alphahot:vhotburst:epsilon_Star	-0.05	0.02
alphahot:alpha_reheat:vhotburst	0.05	0.02

Table 3: The polynomial fit for Wave 1, Output K21

	Estimate	Std. Error
(Intercept)	-1.77	0.01
vhotdisk	-0.34	0.02
alphahot	-0.19	0.03
alpha_reheat	0.11	0.03
vhotburst	-0.14	0.02
stabledisk	-0.03	0.02
I(alpha_reheat^2)	-0.11	0.02
I(vhotdisk^2)	-0.06	0.02
I(alpha_reheat^3)	0.10	0.04
I(stabledisk^2)	-0.03	0.02
I(vhotburst^2)	-0.03	0.02
I(alphahot^3)	0.06	0.04
vhotburst:stabledisk	-0.23	0.02
vhotdisk:alphahot	-0.22	0.02
alphahot:vhotburst	-0.09	0.02
alphahot:stabledisk	-0.10	0.02
vhotdisk:alpha_reheat	0.11	0.02
alpha_reheat:vhotburst	0.09	0.02
alphahot:alpha_reheat	0.07	0.02
alpha_reheat:stabledisk	0.07	0.02
stabledisk:I(vhotdisk^2)	-0.07	0.03
alphahot:I(vhotdisk^2)	-0.06	0.03
alphahot:I(stabledisk^2)	-0.07	0.03
vhotdisk:I(stabledisk^2)	-0.07	0.03
vhotburst:I(alpha_reheat^2)	-0.06	0.03
stabledisk:I(vhotburst^2)	-0.09	0.04
alphahot:I(vhotburst^2)	-0.06	0.04
alphahot:vhotburst:stabledisk	-0.11	0.03
alpha_reheat:vhotburst:stabledisk	0.10	0.03
alphahot:alpha_reheat:stabledisk	0.06	0.03
vhotdisk:alphahot:alpha_reheat	0.04	0.03

Table 4: The polynomial fit for Wave 1, Output K22.25

	Estimate	Std. Error
(Intercept)	-2.63	0.04
alpha_cool	-0.84	0.07
vhotdisk	-0.62	0.05
vhotburst	-0.47	0.06
alpha_reheat	0.41	0.07
stabledisk	-0.17	0.07
I(vhotdisk^2)	-0.21	0.05
I(alpha_cool^2)	-0.19	0.05
I(vhotburst^2)	-0.16	0.05
I(alpha_reheat^2)	-0.14	0.05
I(alpha_cool^3)	0.27	0.09
I(stabledisk^2)	0.10	0.05
I(stabledisk^3)	-0.16	0.09
vhotburst:stabledisk	-0.48	0.04
alpha_cool:stabledisk	-0.47	0.04
alpha_cool:vhotburst	-0.41	0.04
alpha_reheat:stabledisk	0.32	0.04
alpha_cool:vhotdisk	-0.24	0.04
alpha_cool:alpha_reheat	0.28	0.04
vhotburst:alpha_reheat	0.27	0.04
vhotdisk:stabledisk	-0.20	0.04
vhotdisk:alpha_reheat	0.16	0.04
stabledisk:I(vhotburst^2)	-0.27	0.08
alpha_reheat:I(alpha_cool^2)	-0.22	0.08
vhotdisk:I(alpha_reheat^2)	0.24	0.08
vhotdisk:I(alpha_cool^2)	0.19	0.08
alpha_cool:I(stabledisk^2)	0.27	0.08
alpha_cool:I(vhotburst^2)	-0.15	0.08
vhotdisk:I(stabledisk^2)	0.17	0.08
vhotdisk:vhotburst	-0.09	0.04
alpha_reheat:I(alpha_reheat^2)	0.15	0.09
vhotburst:I(vhotburst^2)	0.14	0.09
alpha_reheat:I(vhotdisk^2)	-0.12	0.08
alpha_cool:vhotburst:stabledisk	-0.45	0.07
alpha_cool:vhotdisk:stabledisk	-0.27	0.07
vhotburst:alpha_reheat:stabledisk	0.33	0.07
alpha_cool:vhotburst:alpha_reheat	0.19	0.07
alpha_cool:alpha_reheat:stabledisk	0.23	0.07
vhotdisk:vhotburst:alpha_reheat	-0.19	0.07

Table 5: The polynomial fit for Wave 1, Output K24.75

	Estimate	Std. Error
(Intercept)	-3.70	0.04
alpha_cool	-1.37	0.05
stabledisk	-0.30	0.08
alpha_reheat	0.47	0.08
vhotburst	-0.61	0.07
vhotdisk	-0.74	0.06
I(stabledisk^2)	0.40	0.05
I(stabledisk^3)	-0.43	0.10
I(vhotdisk^2)	-0.24	0.05
I(alpha_cool^2)	0.19	0.05
I(alpha_reheat^2)	-0.18	0.05
I(vhotburst^2)	-0.16	0.05
I(alpha_reheat^3)	0.21	0.10
alpha_cool:stabledisk	-0.63	0.05
stabledisk:vhotdisk	-0.40	0.05
stabledisk:vhotburst	-0.48	0.05
alpha_cool:vhotburst	-0.46	0.04
stabledisk:alpha_reheat	0.39	0.05
alpha_cool:I(stabledisk^2)	0.58	0.09
alpha_cool:vhotdisk	-0.24	0.05
alpha_reheat:vhotburst	0.24	0.04
vhotdisk:I(stabledisk^2)	0.43	0.09
alpha_reheat:vhotdisk	0.19	0.05
vhotdisk:I(alpha_cool^2)	0.30	0.09
alpha_reheat:I(alpha_cool^2)	-0.25	0.09
alpha_cool:I(vhotdisk^2)	0.24	0.09
vhotburst:vhotdisk	-0.11	0.05
vhotburst:I(alpha_cool^2)	0.20	0.09
stabledisk:I(vhotburst^2)	-0.21	0.09
vhotdisk:I(alpha_reheat^2)	0.19	0.09
alpha_cool:alpha_reheat	0.11	0.05
vhotburst:I(vhotburst^2)	0.20	0.10
alpha_reheat:I(vhotdisk^2)	-0.17	0.09
alpha_cool:I(alpha_reheat^2)	0.17	0.09
vhotdisk:I(vhotburst^2)	0.12	0.09
stabledisk:I(alpha_cool^2)	0.12	0.09
alpha_cool:stabledisk:vhotburst	-0.46	0.08
alpha_cool:stabledisk:vhotdisk	-0.34	0.08
stabledisk:alpha_reheat:vhotburst	0.28	0.08
alpha_reheat:vhotburst:vhotdisk	-0.28	0.08
alpha_cool:alpha_reheat:vhotdisk	-0.22	0.08
alpha_cool:stabledisk:alpha_reheat	0.16	0.08
stabledisk:alpha_reheat:vhotdisk	0.15	0.08

Table 6: The polynomial fit for Wave 1, Output K25.75

	Estimate	Std. Error
(Intercept)	-1.32	0.00
vhotdisk	-0.23	0.01
alphahot	-0.16	0.01
alpha_reheat	0.05	0.01
epsilon_Star	-0.01	0.01
vhotburst	-0.03	0.00
I(vhotdisk^2)	-0.08	0.01
I(epsilon_Star^2)	-0.08	0.01
I(epsilon_Star^3)	0.10	0.01
I(alpha_reheat^2)	-0.05	0.01
I(alphahot^2)	-0.04	0.01
I(vhotdisk^3)	0.04	0.01
I(alpha_reheat^3)	0.04	0.01
vhotdisk:alphahot	-0.20	0.01
vhotdisk:alpha_reheat	0.09	0.01
alphahot:alpha_reheat	0.05	0.01
vhotdisk:epsilon_Star	0.04	0.01
alphahot:epsilon_Star	0.03	0.01
vhotdisk:I(epsilon_Star^2)	0.06	0.01
vhotdisk:vhotburst	0.02	0.01
vhotdisk:I(alpha_reheat^2)	-0.04	0.01
vhotdisk:I(alphahot^2)	-0.03	0.01
alphahot:I(alpha_reheat^2)	-0.03	0.01
epsilon_Star:vhotburst	-0.01	0.01
alphahot:I(vhotdisk^2)	-0.02	0.01
alpha_reheat:I(alphahot^2)	0.02	0.01
alphahot:I(epsilon_Star^2)	0.01	0.01
vhotdisk:alphahot:epsilon_Star	0.04	0.01
vhotdisk:alphahot:alpha_reheat	0.03	0.01

Table 7: The polynomial fit for Wave 1, Output bj 17

	Estimate	Std. Error
(Intercept)	-2.85	0.03
alpha_cool	-0.70	0.06
alpha_reheat	0.34	0.03
vhotdisk	-0.27	0.03
vhotburst	-0.24	0.02
epsilon_Star	-0.08	0.06
I(epsilon_Star^2)	0.18	0.04
I(vhotburst^2)	-0.14	0.04
I(alpha_cool^3)	0.24	0.07
I(alpha_reheat^2)	-0.12	0.04
I(vhotdisk^2)	-0.10	0.04
I(epsilon_Star^3)	-0.14	0.07
alpha_reheat:vhotburst	0.22	0.03
alpha_cool:vhotburst	-0.20	0.03
alpha_cool:vhotdisk	-0.16	0.03
vhotburst:epsilon_Star	-0.15	0.03
alpha_cool:alpha_reheat	0.16	0.03
alpha_reheat:vhotdisk	0.14	0.03
vhotdisk:vhotburst	-0.09	0.03
alpha_reheat:I(epsilon_Star^2)	-0.13	0.06
epsilon_Star:I(vhotburst^2)	-0.15	0.07
vhotdisk:I(alpha_reheat^2)	0.13	0.07
alpha_cool:I(vhotdisk^2)	0.12	0.07
alpha_reheat:I(vhotdisk^2)	-0.11	0.06
vhotdisk:epsilon_Star	-0.06	0.03
alpha_cool:I(epsilon_Star^2)	0.13	0.07
epsilon_Star:I(vhotdisk^2)	0.11	0.06
alpha_cool:alpha_reheat:vhotburst	0.11	0.06

Table 8: The polynomial fit for Wave 1, Output bj 21

	Estimate	Std. Error
(Intercept)	-4.07	0.03
alpha_cool	-0.72	0.03
vhotburst	-0.45	0.03
alpha_reheat	0.17	0.05
vhotdisk	-0.56	0.06
stabledisk	0.06	0.06
I(stabledisk^2)	0.35	0.04
I(stabledisk^3)	-0.52	0.08
I(alpha_cool^2)	0.25	0.04
I(alpha_reheat^2)	-0.12	0.04
I(vhotburst^2)	-0.11	0.04
I(alpha_reheat^3)	0.15	0.08
I(vhotdisk^3)	0.16	0.08
vhotdisk:stabledisk	-0.36	0.04
alpha_cool:vhotburst	-0.34	0.03
vhotburst:stabledisk	-0.30	0.04
alpha_cool:I(stabledisk^2)	0.45	0.07
vhotburst:alpha_reheat	0.22	0.03
alpha_cool:stabledisk	-0.22	0.04
alpha_reheat:stabledisk	0.21	0.04
vhotdisk:I(stabledisk^2)	0.45	0.07
alpha_cool:vhotdisk	-0.13	0.04
stabledisk:I(alpha_cool^2)	0.27	0.07
vhotburst:I(alpha_cool^2)	0.20	0.07
vhotdisk:I(alpha_cool^2)	0.24	0.07
alpha_reheat:vhotdisk	0.11	0.04
stabledisk:I(vhotburst^2)	-0.19	0.07
vhotburst:vhotdisk	-0.10	0.04
vhotburst:alpha_reheat:stabledisk	0.26	0.06
alpha_cool:vhotburst:stabledisk	-0.21	0.06
alpha_cool:vhotdisk:stabledisk	-0.10	0.06
vhotburst:alpha_reheat:vhotdisk	-0.09	0.06

Table 9: The polynomial fit for Wave 1, Output bj 22.25